Colorectal Cancer Survival Prediction Using Deep Distribution Based Multiple-Instance Learning

Xingyu Li, Jitendra Jonnagaddala, Min Cen, Hong Zhang, and Steven Xu

1. School of Management, University of Science and Technology of China, Hefei 230026, China
2. School of Population Health, University of New South Wales, Sydney, NSW 2052, Australia
3. Clinical Pharmacology and Quantitative Science, Genmab US, Inc., Princeton, NJ 08540, USA
* Correspondence: jitendra.jonnagaddala@unsw.edu.au (J.J.); zhangh@ustc.edu.cn (H.Z.); sxu@genmab.com (S.X.)

Abstract: Most deep-learning algorithms that use Hematoxylin- and Eosin-stained whole slide images (WSIs) to predict cancer survival incorporate image patches either with the highest scores or a combination of both the highest and lowest scores. In this study, we hypothesize that incorporating wholistic patch information can predict colorectal cancer (CRC) cancer survival more accurately. As such, we developed a distribution-based multiple-instance survival learning algorithm (DeepDisMISL) to validate this hypothesis on two large international CRC WSIs datasets called MCO CRC and TCGA COAD-READ. Our results suggest that combining patches that are scored based on percentile distributions together with the patches that are scored as highest and lowest drastically improves the performance of CRC survival prediction. Including multiple neighborhood instances around each selected distribution location (e.g., percentiles) could further improve the prediction. DeepDisMISL demonstrated superior predictive ability compared to other recently published, state-of-the-art algorithms. Furthermore, DeepDisMISL is interpretable and can assist clinicians in understanding the relationship between cancer morphological phenotypes and a patient’s cancer survival risk.

Keywords: survival analysis; deep learning; multiple instance learning; whole slide images

1. Introduction

Traditional risk stratification in cancer patients is usually based on cancer staging, grading, and molecular and clinical characteristics [1–3]. Tumor grade based on cellular appearance has been shown to be a measure of prognosis and an indicator of how quickly a tumor is likely to grow and spread [4]. However, high inter-observer variability has been a major challenge [5]. With the advances in computer vision, deep learning algorithms have been successfully applied to Hematoxylin- and Eosin- (H&E) stained whole slide images (WSIs) to predict patient outcomes such as overall survival, progression-free survival, time to metastasis, or tumor recurrence. Additionally, WSIs were also used to stratify patients according to their survival risk [6–9].

WSIs are usually split into small image patches to train neural networks to predict an outcome of interest. Due to a large number of image patches for each WSI that can vary from hundreds to thousands, and the unavailability of patch-level labels, it is not feasible to directly build a prediction model at the patch level. Zhu et al. and Wulczyn et al. randomly selected patches to train their deep-learning survival models [10,11]. In addition, unsupervised clustering was utilized to facilitate the identification of the most predictive image patches for survival modeling [6,7,9,12]. Furthermore, tumor/stromal regions are often considered the most predictive. Therefore, tissue segmentation models were used to select tumor/stromal patches for model building [13,14]. Recently, Courtiol et al. proposed MesoNet for survival data which demonstrated that using both the lowest-scored patches and the highest-scoring patches can more accurately predict risk and survival in malignant mesothelioma patients [8].
In this study, we hypothesize that incorporating holistic patch information can predict CRC cancer survival more accurately. Therefore, we developed a distribution-based multiple-instance survival learning algorithm (DeepDisMISL) to validate this hypothesis. DeepDisMISL is also highly interpretable and can assist clinicians in the identification of morphological phenotypes. Using a 5-fold cross-validation on a large Australian colorectal cancer dataset—"MCO CRC"—and external validation on the TCGA COAD-READ dataset, we demonstrate that, by gradually incorporating patches scored and stratified into different percentiles (e.g., the 1st, 5th, 25th, 75th, 95th, and 99th percentiles) to the patches with highest and lowest scores, the predictive performance of CRC survival prediction can be improved dramatically [15,16]. In addition, we systematically compared DeepDisMISL with six different baseline algorithms. The proposed DeepDisMISL demonstrates superior and robust predictive ability compared to all six baseline algorithms, including recently-published state-of-the-art algorithms such as MesoNet and DeepAttnMISL [6,8].

2. Materials and Methods

2.1. Datasets

The analysis used two large-scale datasets. WSIs (40×) were collected from both the MCO CRC and TCGA CRC studies. The MCO CRC dataset was made available through the SREDH Consortium (www.sredhconsortium.org, accessed on 15 January 2021) [15,16] and consists of patients who underwent curative resection for colorectal cancer between 1994 to 2010 in New South Wales, Australia. The Cancer Genome Atlas (TCGA) public dataset includes the TCGA-COAD and TCGA-READ datasets. The MCO CRC dataset (15 January 2021) was used to train the deep learning model while the TCGA COAD and READ dataset (obtained in July 2020) was used for external validation. For patients with more than one WSI available, we randomly selected one slide for each patient. Images with annotation marks or blur were excluded. After exclusion, 1184 WSIs from 1184 patients from the MCO CRC dataset were included in the analysis, whereas 529 WSIs from 529 patients were available from the TCGA database.

2.2. Preprocessing of WSIs

WSIs were first preprocessed to exclude the background area of each image where no tissue was present. The OTSU algorithm was implemented to classify the image into two classes: the foreground (containing the matter) and the background [17]. After background removal, the WSIs were cropped into non-overlapping, fixed-size tiles of images (224 × 224 pixels, 0.5 mpp), which were then color-normalized with Macenko’s method [18]. Depending on the size of the image, the number of tiles varied from a few hundred to 50,000 with a median value of 12,047.

2.3. Feature Extraction

We extracted features using a fine-tuned Xception model, which has been shown to provide improved feature extraction [19,20]. This neural network allowed us to obtain 256 relevant features from each tile. For each WSI (patient), a feature matrix with a dimension of n (number of tiles) × 256 (features) was obtained. We randomly selected 12,000 tiles from the feature matrix of each patient. If the tile number per slide was smaller than 12,000, we would resample the feature matrix up to a size of 12,000.

2.4. DeepDisMISL

\[ L(O_i) = \sum_i \delta_i (-O_i - \log \sum_{j|O_j \geq O_i} \exp(O_j)) \]  

(1)

The log-likelihood function of the Cox loss function used is Equation (1), where \( O_i \) is the risk score, \( \delta_i \) is the censoring variable (\( \delta = 0 \), death has not been observed). The neural network minimized the negative log partial likelihood function and produced maximum partial likelihood estimates.
We started with top/bottom instances (i.e., minimum, and maximum patch scores, or 0 and 100th percentiles), and incrementally added additional instances at different percentiles to the top/bottom instances. In total, seven different scenarios were investigated:

(Scenario 1) [0, 100%],
(Scenario 2) [0, 0.1%, 99.9%, 100%],
(Scenario 3) [0, 0.1%, 1%, 99.9%, 100%],
(Scenario 4) [0, 0.1%, 1%, 5%, 95%, 99%, 99.9%, 100%],
(Scenario 5) [0, 0.1%, 1%, 5%, 10%, 90%, 95%, 99%, 99.9%, 100%],
(Scenario 6) [0, 0.1%, 1%, 5%, 10%, 25%, 75%, 90%, 95%, 99%, 99.9%, 100%], and
(Scenario 7) [0, 0.1%, 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%, 99.9%, 100%].

Zhu et al. showed that increasing the number of top instances could improve model performance [7]. We also evaluated whether the number of neighborhood instances around each percentile had an impact on the model’s predictive ability. In this experiment, we took the model with the most complete distribution (Scenario #7) and explored different numbers of neighborhood instances at each percentile (1, 3, 5, and 7).

2.5. Baseline Algorithms

We compared our algorithm with the following baseline algorithms (Figure 1): MesoNet, Meanpooling and Maxpooling (Top 1 and Top 10 instances), MeanFeaturePool, and DeepAttnMISL [6,8,21]. In common with DeepDisMISL, for MesoNet we used two 1-D convolution layers to aggregate all local feature descriptors of a tile into a global feature vector. Then, the largest and smallest 10 scores from the global feature vector were selected as the features for the classifier in the prediction layer. Similarly, Meanpooling and Maxpooling (Top 1 instance) used the mean score and max score, respectively, from the global feature vector generated by 1-D convolution layers as the features for the classifier in the prediction layer. Maxpooling (Top 10 instance) used the largest 10 scores from the global feature vector as the features for the classifier in the prediction layer. For Meanpool (LASSO Cox), the features values for all tiles (12,000) from a feature matrix (12,000 × 256) for each WSI were averaged to obtain an average feature vector (1 × 256) for each WSI or patient [22]. LASSO Cox implemented in the R package glmnet was used to fit the survival data. The C-index values were obtained from a 5-fold cross-validation. Both DeepAttnMISL and our proposed DeepDisMISL used the MCO CRC dataset for cross-validation. Therefore, the concordance index (C-index) values for DeepAttnMISL from Ref [6] were obtained and compared directly with DeepDisMISL [6]. We further compared the performance of DeepDisMISL with the other baseline algorithms by examination of their ability for risk stratification. For each model, the median risk score in the training set was calculated and then applied as a threshold to stratify each patient into the high-risk or low-risk group.

Figure 1 shows the proposed DeepDisMISL algorithm. We incorporated multiple-instance learning (MIL) before predicting on the unlabeled tiles. WSIs were considered the bags of MIL (each WSI can be considered as a bag, and patches of the WSI can be seen as instances), for which supervision was provided only for the bag at individual patient level (i.e., survival time t and survival status δ for a patient), and the individual label of the instances contained in the bags (i.e., tiles) were not available. Two convolution, one-dimensional (1-D) layers with ReLU activation were used to aggregate all the local features of the same tile into a global feature and devise a score (the element in the vector of 12,000 × 1, each element can be seen as a score, Figure 1 and Table 1) for each tile. The scores at the selected percentiles (instances) of the global feature vector were the input features for the next prediction layer, which consisted of a two-layer multi-layer perceptron (MLP) classifier with two fully connected layers (128 and 64 neurons) and a ReLU activation for predicting survival risk with a Cox loss function to deal with the censored survival data.
Figure 1. The proposed DeepDisMISL algorithm. (a) DeepDisMISL with single instance at each percentile; (b) DeepDisMISL with multiple instances at each percentile; (c) DeepDisMISL with only top and bottom instances.

Table 1. The structure of the DeepDisMISL algorithm.

| Layer               | Input        | Output Size          |
|---------------------|--------------|----------------------|
| 1D convolution layer| 12,000 × 256 | 12,000 × 128         |
| 1D convolution layer| 12,000 × 128 |                      |
| Aggregating layer   | 12,000 × 1   | \(m_i \times 1\)     |
| Fully-Con.          | \(m_i \times 1\) | 128 × 1              |
| Fully-Con.          | 128 × 1      | \(O_i\)              |

\(m_i\): the number of percentile points; \(O_i\): risk score.

2.6. Annotation of WSI Patches

To interpret the model, we annotated the patches near the different percentiles of the patch scores. Kather et al. developed a deep-learning classifier to classify CRC image tiles into eight tissue types: adipose tissue (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM) [23]. We used the pathologist-annotated NCT-CRC-HE-100K and CRC-VAL-HE-7K image sets provided by Kather et al. to train and validate, respectively, a similar tissue-type classifier. We downloaded the Xception model from Keras and fine-tuned the model using the NCT-CRC-HE-100K image set to develop the tissue-type classifier [20]. The overall accuracy of the tissue-type classification model was 99% based on the training dataset, NCT-CRC-HE-100K, and 94.4% based on...
the validation image set, CRC-VAL-HE-7K (Supplementary Figure S1). The tissue type of each image tile from the TCGA dataset was predicted using the fine-tuned Xception-based tissue-type classifier.

2.7. Evaluation

All models were trained using 5-fold cross-validation on the MCO CRC dataset. In each fold, 80% of the data were used for model training and 20% of the data were used for model validation. For training, we used Adam optimization with a grid search strategy. The training process monitored the loss on the MCO validation dataset and was designed to stop if the loss increased. C-index is widely used in deep-learning models for censored survival data [6,8,24,25]. We evaluated model performance with the C-index in the survival prediction task. The TCGA data served as the independent, external validation dataset.

3. Results

3.1. Evaluation of Probability Distribution-Based Patch Selection

Figure 2a shows that, based on the 5-fold cross-validation using the MCO CRC dataset, there is an obvious trend where the more percentiles utilized, the higher the C-index. With the top/bottom instances (See Figure 1c), the average C-index was 0.611 (range: 0.58–0.630). Adding two more percentiles at 0.1% and 99.9% (Scenario #2) improved the average C-index to 0.62 (0.59–0.638). As expected, Scenario #7 with the most complete distribution information (0, 0.1%, 1%, 5%, 10%, 90%, 25%, 50%, 75%, 95%, 99%, 99.9%, and 100%) produced the best predictive performance with an average C-index of 0.638 (0.626–0.66). The increasing trend in C-index suggests that the complete distribution of the patch scores carries richer information on the WSI than the only top and bottom instances (Figure 3).

Similarly, the external validation using an independent TCGA COAD-READ dataset demonstrated a similar trend (Figure 2b). In general, the models with more complete distribution (i.e., more middle-scoring patches at different distribution percentiles in addition to the highest/lowest scoring patches) provide a higher C-index compared to those with less complete distribution (e.g., only the highest and lowest scoring patches). As expected, the external validation with the independent dataset produced more heterogeneous results and a lower C-index compared to the cross-validation, since the TCGA dataset does not completely resemble the MCO CRC dataset. Nevertheless, the external validation with the TCGA dataset still preserved the overall trend wherein a more complete distribution of the patch scores provided better predictive ability than the extreme top/bottom instances alone.

3.2. Multiple Neighborhood Instances vs. Single Instance at Each Percentile

Figure 3 clearly shows that multiple instances outperformed the single instance at each percentile. For the cross-validation using the MCO CRC dataset, the average C-index was improved from 0.640 to 0.645 when the number of instances at each percentile increased from one to three (Figure 3a). The average C-index increased to 0.647 with five neighborhood instances, and the improvement appears to level off with more neighborhood instances (i.e., seven). A similar pattern was observed with the independent TCGA dataset. The average C-index plateaued (0.580) when the number of instances at each percentile increased to three. Therefore, consistent with Zhu et al.’s findings, multiple neighborhood instances at each percentile can improve the model performance [7]. However, experiments may be needed to determine the optimal number of instances for different tasks and models.

3.3. Comparison with Baseline Algorithms

Our proposed DeepDistMISL algorithm demonstrated superior predictive ability compared to all baseline algorithms (Figure 4). Compared to MesoNet, DeepDistMISL provided an additional 6.3% and 2.8% improvement of mean C-index in the 5-fold cross-validation and external validation, respectively. In addition, DeepDisMISL also outperformed the most recently-published, state-of-the-art algorithm, DeepAttnMISL for the MCO CRC dataset.
The mean C-index of DeepDisMISL was markedly higher than that of DeepAttnMISL (0.647 vs. 0.606) for the MCO CRC dataset. The superiority of the proposed DeepDistMISL algorithm in both cross-validation and external validation indicates the robustness of this algorithm and highlights the importance of using a complete distribution of patch scores in predictive models. Furthermore, Maxpooling (both top 1 and top 10 instances) had the worst performance compared to the other approaches in both cross-validation and external validation. Similar to the finding in Courtiol et al., although MeanFeaturePool provided better performance than MesoNet in cross-validation, MeanFeaturePool seemed less robust and had a lower C-index in the external validation (Figure 4b) compared to MesoNet [8]. It is interesting to notice that the simple Meanpooling approach had the second-best performance. Meanpooling outperformed all the other baseline approaches in the internal cross-validation using the MCO CRC dataset and maintained its performance in the external validation (i.e., provided a similar C-index to MesoNet and outperformed other baseline algorithms) [8]. It should be mentioned that the results of DeepAttnMISL were obtained directly from the previous publication, and no external validation was conducted for DeepAttnMISL [6].

![Five fold cross validation on the MCO Dataset](figure)

![External validation on the TCGA COAD-READ dataset](figure)

**Figure 2.** Performance for models with different percentiles (single tile at each percentile) for (a) the MCO CRC dataset (internal validation) and (b) the TCGA dataset (external testing). Black solid dots = mean C-index value from 5-fold cross-validation experiments; Red solid dots = individual C-index values; Error bar = standard deviation.
Figure 3. Effects of the number of neighborhood tiles at each percentile on model performance for (a) the MCO CRC dataset (internal validation) and (b) the TCGA dataset (external testing). Solid line = smoothing curve; shaded area = standard deviation around the smoothing line; Black solid dots = mean C-index value from 5-fold cross-validation experiments; Red solid dots = individual C-index values; Error bar = standard deviation.

3.4. Risk Stratification

Figure 5 shows that DeepDisMISL provided the best risk stratification for both the MCO and TCGA populations. Among all the studied deep learning algorithms, DeepDisMISL provides the most statistically-significant separation of the survival curves between the high- and low-risk groups in both the MCO and TCGA populations ($p < 0.0001$ in MCO and $p = 0.01$ in TCGA). That is, DeepDisMISL identified that the high-risk subgroup presented significantly worse overall survival compared to the low-risk subgroup. MesoNet-identified risk groups also showed apparent separation for survival, but with slightly larger $p$ values ($p = 0.001$ in MCO and $p = 0.02$ in TCGA) compared to those based on DeepDisMISL. However, the other baseline algorithms (Meanpooling, Maxpooling with top 1 instance, Maxpooling with top 10 instances, and MeanFeaturePool) only provided clear risk stratification in the MCO CRC
(training) dataset. In the TCGA COAD READ (external validation) dataset, no statistically significant separation was observed for these baseline algorithms.

Figure 4. Comparison of different baseline algorithms using (a) the MCO CRC dataset (internal validation) and (b) the TCGA dataset (external testing).
DeepDisMISL

MesoNet

Meanpooling

Figure 5. Kaplan–Meier plots comparing different algorithms using the MCO CRC dataset (internal validation) and TCGA dataset (external testing). For each algorithm, the median risk score in the training set was calculated and then applied as a threshold to stratify each patient into the high-risk or low-risk group.
3.5. Interpreting DeepDisMISL

Figure 6 shows the relationship between the risk and the percentiles of the tile scores. Every percentile appears correlated with risk, i.e., a positive relationship was observed between risk and percentile of tile scores for 0, 1/10th, 1st, 5th, 10th, and 25th percentiles whereas a negative relationship was observed for 50th, 75th, 90th, 99th, 99.9th, and 100th percentiles. The strong correlation between risk and individual tile score percentiles may explain the decent performance of some published algorithms such as Maxpooling and Meanpooling. In addition, the tile scores at individual percentiles may carry different information regarding the survival risk. Therefore, combining certain individual percentiles (such as in MesoNet, top/bottom 10 instances were used as prediction features) has been shown to provide better predictive performance [8]. This may also explain the excellent performance of DeepDisMISL and illustrate the importance of utilizing the information of the entire distribution (i.e., combining multiple percentiles). Figure 7 shows the representative tiles near the percentiles. It is interesting to notice that at lower percentiles (e.g., 0–75%), the tiles primarily included tumor cells. Conversely, at the higher percentiles (e.g., 90%, 95, 99%, 99.9%, and 100%), muscle appears to be the predominant tissue type. This may partly explain why at lower percentiles there appears to be a positive relationship between risk and the tile scores, whereas a reversed, negative relationship between risk and tile scores was observed at higher percentiles (Figure 6). It is intuitive that the tumor patches are related to higher risk, while normal muscle patches may represent lower risk. This suggests that our algorithm is well interpretable and can help to reveal the relationship between morphological phenotypes and a patient’s risk.

![Figure 6](imageURL)

**Figure 6.** Relationship between the average tile score at each percentile and the risk. Patients are stratified into 10 different risk groups.

3.6. Attention Aggregation

We attempted to use an attention mechanism in addition to the percentile structure on multiple neighborhood instances i.e., we assigned the same weight to each percentile location. However, as the results in Table 2 show, the attention mechanism did not improve the predictive ability. The C-index on the MCO dataset with cross-validation from the attention-based model was 0.627, while the C-index using the external validation dataset TCGA was 0.566. It is worth noting that due to the large number of parameters required for attention mechanism-based models, overfitting may be a challenge and prevent further improvement in the prediction.
0 STR TUM TUM 0.1% STR TUM TUM

1% TUM TUM TUM 5% STR ADI TUM

10% NORM MUS STR 25% TUM TUM TUM

50% STR TUM TUM 75% TUM TUM TUM

90% MUS MUS MUS 95% MUS MUS MUS

99% MUS MUS MUS 99.9% MUS MUS MUS

100% STR STR TUM

**Figure 7.** Visualization of the spatial locations and morphologic features of tiles near different percentiles in a whole slide image. The thirteen different colors represent thirteen percentile locations. Three neighborhood tiles at each percentile are selected for visualization, and the tissue type of each tile is predicted by a published CRC tissue classifier.
Table 2. Performance with and without an attention mechanism.

|                      | MCO CRC Dataset-Internal Validation | TCGA COAD-READ Dataset-External Validation |
|----------------------|-------------------------------------|--------------------------------------------|
|                      | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | No attention mechanism | With attention mechanism | No attention mechanism | With attention mechanism |
| No attention mechanism | 0.66   | 0.64   | 0.65   | 0.65   | 0.65   | 0.61                     | 0.65                     | 0.58                     | 0.59                     |
| With attention mechanism | 0.67   | 0.65   | 0.62   | 0.61   | 0.6    | 0.55                     | 0.6                      | 0.56                     | 0.54                     |

4. Discussion

Risk stratification for cancer patients is currently largely based on cancer staging, mutation/molecular subtyping, and clinical features. Recently, outcome prediction algorithms based on histopathology images using deep learning have been proposed to stratify patients [7,10,13,14,23,26]. Although different approaches have been proposed to develop deep learning survival models, identification of the most predictive image patches is an attractive strategy and has been an active area of research. Patches with the highest score are often considered to carry the most predictive value. In addition, Courtiol et al. trained a deep learning model (MesoNet) and demonstrated that adding the patches with the lowest scores to the patches with the highest scores can provide an excellent prediction of survival in patients with mesothelioma [8].

In this study, we proposed DeepDisMISL, a patch-score distribution-based multiple-instance survival learning algorithm, and demonstrated that other patches also carry important information required to predict survival. We believe this is due to the fact that patches don’t convey the complete clinical scenario of a tumor. As such, in common with clinicians, algorithms need to leverage all the information available to form a clinical impression. By incrementally adding additional patch scores at different percentiles (e.g., 1st, 5th, 25th, 75th, 95th, and 99th percentiles) to the highest and lowest scoring patches, the predictive performance of DeepDisMISL for survival prediction in colorectal cancer was improved dramatically.

We also systematically compared our proposed DeepDisMISL with six existing models: MesoNet, DeepAttnMISL, Meanpooling, Maxpooling (Top 1 instance), Maxpooling (Top 10 instance), and MeanFeaturePool [6,8]. DeepDisMISL was not only superior to MesoNet but also outperformed the most recently-published state-of-the-art DeepDisMISL. The mean C-index of our proposed DeepDisMISL (0.647) was markedly higher than that of DeepAttnMISL (C-index = 0.606) for the MCO CRC dataset [6]. Compared to MesoNet, DeepDisMISL provided a 6.3% and 2.8% improvement of mean C-index in the 5-fold cross-validation and external validation, respectively.

The interpretability of deep-learning models is critical and can help the understanding of the underlying pathology and inform future directions for model improvements. The MesoNet identified image patches were highly interpretable i.e., the high-risk patches were mainly located in stroma regions for patients with mesothelioma. DeepDisMISL was also highly interpretable and revealed a positive relationship between tumor tissues and the risk of death, and a negative relationship between normal muscle tissues and the risk of death. This suggests that DeepDisMISL can help to detect the predictive morphological phenotypes.

It is known that, without external validation, deep-learning models are prone to a high risk of bias due to batch effects. As such, we validated and compared our model to other state-of-the-art algorithms using not only 5-fold internal cross-validation but also externally on an independent TCGA dataset. Both validations showed that the proposed DeepDisMISL provided superior performance over all baseline algorithms, indicative of the robustness of our findings. Further evaluation and applications of DeepDisMISL in other types of cancers and different populations of colorectal cancer patients are warranted.
Our method also has limitations. First, although the characteristics of patients from the MCO (model training) and TCGA cohorts (model validation) appear similar (Table 3), selection bias (e.g., the MCO and TCGA cohorts may contain different patient populations since these are not randomized studies) cannot be ruled out. In addition, the follow-up duration of the MCO cohort (a maximum of 60 months) was significantly shorter than the TCGA cohort (a maximum of 150 months). A randomized, prospective study in the future may provide a better evaluation of model performance in different cohorts. In addition, despite the popularity of Cox loss in deep-learning survival modeling, alternatives such as Uno loss, log-rank loss, the exponential lower bound on the C-index, and multitask classification using cross-entropy may provide better performance [11,27]. Future research may be warranted to further explore alternative loss functions. Finally, our preliminary analysis showed that, compared to ReLU, alternative activation functions (e.g., SeLU) may perform better (data not shown). Further improvement of our model may be achieved by using SeLU in the future [28].

Table 3. The comparison of MCO and TCGA datasets.

|                      | MCO Cohort (n = 1184) | TCGA Cohort (n = 530) |
|----------------------|-----------------------|-----------------------|
| Death                | 379 (32.0%)           | 94 (17.7%)            |
| Age, years (median, range) | 69 (24–99)  | 67 (31–90)            |
| Sex                  |                       |                       |
| Female               | 480 (40.5%)           | 233 (44.0%)           |
| Male                 | 591 (49.9%)           | 263 (49.6%)           |
| Missing              | 113 (9.6%)            | 34 (6.4%)             |
| Stage                |                       |                       |
| Stage I-II           | 601 (52.4%)           | 256 (48.3%)           |
| Stage III-IV         | 468 (40.8%)           | 221 (41.7%)           |
| Missing              | 115 (6.8%)            | 53 (10.0%)            |
| MSI                  |                       |                       |
| MSI-H                | 929 (78.5%)           | 329 (62.1%)           |
| MSS                  | 142 (12.0%)           | 58 (10.9%)            |
| Missing              | 113 (9.5%)            | 143 (27.0%)           |

5. Conclusions

We developed DeepDisMISL, a novel distribution-based multiple-instance survival learning algorithm, to validate our hypothesis that incorporating holistic patch information within a WSI can predict CRC cancer survival more accurately. Instead of using just the patches with the highest and lowest score, we also used patches that were scored based on the percentile distributions. We demonstrated that this approach can dramatically improve the prediction of CRC patient survival outcomes. Including multiple neighborhood instances around each selected distribution location (e.g., percentiles) can further improve the predictive performance. When compared against the six state-of-the-art baseline algorithms, DeepDisMISL demonstrated better prediction performance and more accurate risk stratification for overall survival on both the MCO CRC and TCGA COAD-READ datasets. DeepDisMISL is highly interpretable with the ability to reveal the relationships and interdependencies between morphological phenotypes and a patient’s cancer prognosis risk.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/e24111669/s1, Figure S1. The overall accuracies of the tissue-type
classification model were 99% on (a) the training dataset NCT-CRC-HE-100K and 94.4% on (b) the validation image set xxx, respectively.

**Author Contributions:** S.X., X.L. and H.Z. contributed to design of the research; J.J., X.L. and S.X. contributed to data acquisition; X.L., S.X. and M.C. contributed to data analysis. X.L., S.X., J.J. and H.Z. contributed to data interpretation. X.L., S.X., J.J. and H.Z. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (No. 12171451, 72091212), Anhui Center for Applied Mathematics, Australian National Health and Medical Research Council (No. GNT1192469), Google Cloud Research (award# GCP19980904).

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The MCO CRC dataset was made available through the SREDH Consortium (www.sredhconsortium.org, accessed on 15 January 2021).

**Acknowledgments:** The research of Xingyu Li, Min Cen, and Hong Zhang are partially supported by the National Natural Science Foundation of China (No. 12171451, 72091212) and Anhui Center for Applied Mathematics. Jitendra Jonnagaddala is funded by the Australian National Health and Medical Research Council (No. GNT1192469). Jitendra Jonnagaddala also acknowledges the funding support received through the Research Technology Services at UNSW Sydney, Google Cloud Research (award# GCP19980904) and NVIDIA Academic Hardware grant programs. We also would like to thank the SREDH Biometrics working group for the assistance with the MCO CRC dataset access.

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

**References**

1. Preisser, F.; Cooperberg, M.R.; Crook, J.; Feng, F.; Graefen, M.; Karakiewicz, P.I.; Klotz, L.; Montironi, R.; Nguyen, P.L.; D’Amico, A.V. Intermediate-risk Prostate Cancer: Stratification and Management. *Eur. Urol. Oncol.* 2020, 3, 270–280. [CrossRef] [PubMed]

2. Yu, F.; Quan, F.; Xu, J.; Zhang, Y.; Xie, Y.; Zhang, J.; Lan, Y.; Yuan, H.; Zhang, H.; Cheng, S.; et al. Breast cancer prognosis signature: Linking risk stratification to disease subtypes. *Brief. Bioinform.* 2019, 20, 2130–2140. [CrossRef] [PubMed]

3. Clifford, G.M.; Alberts, C.J. Molecular Risk Stratification for Anal Cancer Prevention. *Clin. Infect. Dis.* 2021, 72, 2164–2166. [CrossRef]

4. Rakha, E.A.; Reis-Filho, J.S.; Baehner, F.; Dabbs, D.J.; Decker, T.; Eusebi, V.; Fox, S.B.; Ichihara, S.; Jacquemier, J.; Lakhani, S.R.; et al. Breast cancer prognostic classification in the molecular era: The role of histological grade. *Breast Cancer Res.* 2010, 12, 207. [CrossRef] [PubMed]

5. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; Chen, R.J.; Barbieri, M.; Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slate images. *Nat. Biomed. Eng.* 2021, 5, 555–570. [CrossRef]

6. Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* 2020, 65, 101789. [CrossRef]

7. Zhu, X.; Yao, J.; Zhu, F.; Huang, J. WSISA: Making Survival Prediction from Whole Slide Histopathological Images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6855–6863.

8. Courtiol, P.; Maussion, C.; Moarri, M.; Pronier, E.; Pilcer, S.; Sefta, M.; Manceron, P.; Toldo, S.; Zaslavskiy, M.; Le Stang, N.; et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* 2019, 25, 1519–1525. [CrossRef] [PubMed]

9. Muhammad, H.; Xie, C.; Sigel, C.S.; Doukas, M.; Alpert, L.; Jarnagin, W.R.; Simpson, A.; Fuchs, T.J. EPIC-Survival: End-to-end Part Inferred Clustering for Survival Analysis, Featuring Prognostic Stratification Boosting. *arXiv* 2021, arXiv:2101.11085.

10. Zhu, X.; Yao, J.; Huang, J. Deep convolutional neural network for survival analysis with pathological images. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 544–547.

11. Wulczyn, E.; Steiner, D.F.; Moran, M.; Plass, M.; Reifs, R.; Tan, F.; Flament-Auvigne, I.; Brown, T.; Regitnig, P.; Chen, P.C.; et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* 2021, 4, 71. [CrossRef]

12. Yue, X.; Dimitriou, N.; Arandjelovic, O. Colorectal Cancer Outcome Prediction from H&E Whole Slide Images using Machine Learning and Automatically Inferred Phenotype Profiles. *arXiv* 2019, arXiv:1902.03582.

13. Yamashita, R.; Long, J.; Longacre, T.; Feng, L.; Berry, G.; Martin, B.; Higgins, J.; Rubin, D.L.; Shen, J. Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study. *Lancet Oncol.* 2021, 22, 132–141. [CrossRef]
14. Klimov, S.; Xue, Y.; Gertych, A.; Graham, R.P.; Jiang, Y.; Bhattarai, S.; Pandol, S.J.; Rakha, E.A.; Reid, M.D.; Aneja, R. Predicting Metastasis Risk in Pancreatic Neuroendocrine Tumors Using Deep Learning Image Analysis. *Front. Oncol.* 2020, 10, 593211. [CrossRef] [PubMed]

15. Jonnagaddala, J.; Croucher, J.L.; Jue, T.R.; Meagher, N.S.; Caruso, L.; Ward, R.; Hawkins, N.J. Integration and Analysis of Heterogeneous Colorectal Cancer Data for Translational Research. *Stud. Health Technol. Inform.* 2016, 225, 387–391. [PubMed]

16. Ward, R.L.; Hawkins, N.J. *Molecular and Cellular Oncology (MCO) Study Whole Slide Image Collection*; UNSW: Sydney, Australia, 2015. [CrossRef]

17. Otsu, N. A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* 1979, 9, 62–66. [CrossRef]

18. Macenko, M.; Niethammer, M.; Marron, J.S.; Borland, D.; Woosley, J.T.; Xiaojun, G.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009; pp. 1107–1110.

19. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv* 2016, arXiv:1610.02357.

20. Macenko, M.; Niethammer, M.; Marron, J.S.; Borland, D.; Woosley, J.T.; Xiaojun, G.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009; pp. 1107–1110.

21. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* 1997, 16, 385–395. [CrossRef]

22. Kather, J.; Krisam, J.; Charoentong, P.; Luedde, T.; Herpel, E.; Weis, C.-A.; Gaiser, T.; Marx, A.; Valous, N.; Ferber, D.; et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* 2019, 16, e1002730. [CrossRef]

23. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* 2016, arXiv:1606.08415.