A Novel and Efficient Digital Pathology Classifier for Predicting Cancer Biomarkers Using Sequencer Architecture

Min Cen,* Xingyu Li,† Bangwei Guo,* Jitendra Jonnagaddala,‡ Hong Zhang,§ and Xu Steven Xu

From the School of Data Science,* University of Science and Technology of China, Hefei, China; the Department of Statistics and Finance,† School of Management, University of Science and Technology of China, Hefei, China; the School of Population Health,‡ University of New South Wales, Sydney, New South Wales, Australia; and Clinical Pharmacology and Quantitative Science,§ Genmab Inc., Princeton, New Jersey

In digital pathology tasks, transformers have achieved state-of-the-art results, surpassing convolutional neural networks (CNNs). However, transformers are usually complex and resource intensive. This study developed a novel and efficient digital pathology classifier called DPSeq to predict cancer biomarkers through fine-tuning a sequencer architecture integrating horizontal and vertical bidirectional long short-term memory networks. Using hematoxylin and eosin–stained histopathologic images of colorectal cancer from two international data sets (The Cancer Genome Atlas and Molecular and Cellular Oncology), the predictive performance of DPSeq was evaluated in a series of experiments. DPSeq demonstrated exceptional performance for predicting key biomarkers in colorectal cancer (micromutator instability status, hypermutation, CpG island methylator phenotype status, BRAF mutation, TP53 mutation, and chromosomal instability), outperforming most published state-of-the-art classifiers in a within-cohort internal validation and a cross-cohort external validation. In addition, under the same experimental conditions using the same set of training and testing data sets, DPSeq surpassed four CNNs (ResNet18, ResNet50, MobileNetV2, and EfficientNet) and two transformer (Vision Transformer and Swin Transformer) models, achieving the highest area under the receiver operating characteristic curve and area under the precision-recall curve values in predicting micromutator instability status, BRAF mutation, and CpG island methylator phenotype status. Furthermore, DPSeq required less time for both training and prediction because of its simple architecture. Therefore, DPSeq appears to be the preferred choice over transformer and CNN models for predicting cancer biomarkers. (Am J Pathol 2023, 193: 2122–2132; https://doi.org/10.1016/j.ajpath.2023.09.006)
being less computationally demanding and resource intensive.

The success of transformers is thought to stem from the self-attention mechanism’s ability to capture long-range dependencies. Although long short-term memory networks can also model long-term dependencies and prevent gradient vanishing through successive time steps, their use in digital pathology is relatively limited and mainly employed for segmentation tasks. Recently, Tatsunami and Taki proposed a sequencer that uses a two-dimensional Bi-directional Long Short-Term Memory (BiLSTM2D) network in ImageNet classification tasks for regular images.

The objective of this research was to develop an efficient digital pathology classifier (DPSeq) by fine-tuning the BiLSTM2D network from sequencer to achieve state-of-the-art predictive performance for critical molecular pathways and gene mutations [ie, MSI, hypermutation, chromosomal instability (CING), Braf mutation, and CpG island methylator phenotype (CIMP)] in colorectal cancer (CRC) using hematoxylin and eosin (H&E)—stained whole-slide images (WSIs). This study provides valuable insights into the comparative performance of DPSeq versus popular transformer (ViT and Swin) and CNN models (ResNet18, MobileNetV2, and EfficientNet) in predicting pathology tasks.

Materials and Methods

The DPSeq was developed to predict cancer biomarkers using WSIs. As illustrated in Figure 1, the classifier development process involved four steps: i) image preprocessing and tile selection, ii) fine-tuning a pretrained sequencer model (trained on the ImageNet data set) using pathology images to build the DPSeq, iii) using the DPSeq to classify tile-level biomarkers, and iv) predicting patient-level biomarkers by aggregating the tile-level predictions. DPSeq performance was assessed by using it to predict MSI, Braf mutation, CIMP, and other key biomarkers for CRC.

Data Sets

Histopathologic WSIs stained with H&E were acquired from two data sets related to CRC. The first data set, named MCO-CRC,14,15 was obtained from the Molecular and Cellular Oncology (MCO) and consists of patients who underwent curative resection for CRC between 1994 and 2010 in New South Wales, Australia. This data set is accessible through the Secure Research Environment for Digital Health (SREDH) Consortium (https://www.sredhconsortium.org, last accessed April 26, 2023), and MSI, Braf mutation, and CIMP ground truth labels were available in this data set. The second data set, TCGA-CRC-DX (https://portal.gdc.cancer.gov, last accessed September 2, 2023), contains formalin-fixed, paraffin-embedded WSIs from two studies of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) conducted by The Cancer Genome Atlas (TCGA), namely TCGA-COAD and TCGA-READ. In addition to the MSI, CIMP, and Braf mutation that were present in the MCO-CRC data set, the TCGA-CRC-DX data set also had hypermutation, CING, and Tp53 mutation information available. The code for this study is available (https://github.com/mimic0127/DPSeq, last accessed May 21, 2023). Numbers of whole-slide images of patients with CRC in these two data sets are listed in Supplemental Table S1. Access to the TCGA and MCO data sets was provided through data usage and ethics agreements. All patients in the TCGA and MCO data sets provided written informed consent.

To develop the tissue classifier and refine the DPSeq, two public data sets that had been annotated by pathologists (these two data sets can be downloaded from https://zenodo.org/record/1214456#.ZC6beOxBxRE, last accessed September 2, 2023) were retrieved. These data sets are referred to as NCT-CRC-HE-100K and CRC-VAL-HE-7K, and they comprise 100,000 and 7180 H&E-stained tiles, respectively. Each tile measures 224 × 224 pixels at 0.5 µm per pixel and has been color normalized via Macenko method.16 The tiles were annotated with nine different tissue types, namely adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium.

Image Preprocessing and Tile Selection

The Laleh et al approach was adopted for processing whole-slide images. All slides from the MCO-CRC and TCGA-CRC-DX data sets were segmented into nonoverlapping tiles of 512 × 512 pixels at 0.5 µm per pixel. The tiles were then color normalized using Macenko method and resized to 224 × 224 pixels to match the input requirement of the deep learning models. A tissue classifier, which was trained and tested on the NCT-CRC-HE-100K and CRC-VAL-HE-7K data sets, respectively,7 was employed to identify tumor tiles. Up to 500 randomly chosen tumor tiles from each patient were used as inputs for the subsequent DPSeq.

DPSeq

DPSeq’s underlying structure was developed using the sequencer framework.10 The primary element of this network is the BiLSTM2D layer, which can integrate both horizontal and vertical patch data within a tile.

Defining the input of BiLSTM2D as \( \textbf{I}_w \in \mathbb{R}^{H \times W \times C} \), the input was split into vertical patch sequences \( \{I_w \}_{W=1}^W \) and horizontal patch sequences \( \{I_h \}_{H=1}^H \). For a fixed \( w \in \{1, W\} \cap \mathbb{Z} \), a vertical patch sequence \( I_{w_0} \) was entered into a bidirectional long short-term memory (BiLSTM) (BiLSTMmer) to extract the
vertical hidden feature: \( H_{\text{vertical}}^{w} = \text{BiLSTM}_{\text{ver}}(I_{w:m}) \). A series of vertical hidden features: \( \{ H_{\text{vertical}}^{w} \in \mathbb{R}^{H \times 2D} \}_{w=1}^{W} \) were obtained where \( D \) is the dimension of hidden layer of \( \text{BiLSTM}_{\text{ver}} \). The weights of \( \text{BiLSTM}_{\text{ver}} \) for all vertical patch sequences were shared. Similarly, a series of horizontal hidden features \( \{ H_{\text{horizontal}}^{h} \in \mathbb{R}^{W \times 2D} \}_{h=1}^{H} \) were obtained by \( H_{\text{horizontal}}^{h} = \text{BiLSTM}_{\text{hor}}(I_{h:.}) \), for all \( h \in [1, H] \cap \mathbb{Z} \).

All vertical hidden features were concentrated to matrix \( H_{\text{vertical}}^{all} \in \mathbb{R}^{W \times H \times 2D} \) and all the horizontal hidden features to matrix \( H_{\text{horizontal}}^{all} \in \mathbb{R}^{W \times H \times 2D} \). The hidden feature for \( I \) was obtained by concatenating \( H_{\text{vertical}}^{all} \) and \( H_{\text{horizontal}}^{all} \) to \( H_{\text{all}} \in \mathbb{R}^{W \times H \times 4D} \). The output of the BiLSTM2D was obtained by a channel fusion with point-wise full connection:

\[
\hat{I} = \text{PointWiseFC}(H_{\text{all}}) = W_{\text{all}}H_{\text{all}} + b_{\text{all}} \in \mathbb{R}^{W \times H \times C}.
\]

As shown in tile-label predictor part of Figure 1, the input tiles had dimensions of 224 × 224 pixels, which were segmented into a total of 32 × 32 smaller patches with a size of 7 × 7 pixels. These patches were processed through a sequence of four stages of sequencer blocks, with each stage containing a different number of blocks (four, three, eight, and four blocks, respectively). The sequencer blocks replaced multihead attention in transformer blocks with BiLSTM2D to improve the efficiency of the classifier. Following layer normalization and average pooling at the end of the sequencer blocks, a three-layer multilayer perceptron with rectified linear unit (ReLU) and dropout layers (384-256-32) was added to extract molecular information from histopathologic images. Finally, a classification layer was added to the top of the network for accurate classification.

To predict biomarkers from histopathologic images, the DPSeq classifier was fine-tuned for multiclass tissue classification. DPSeq parameters were used before the average pooling layer with the parameters of a sequencer model pretrained on ImageNet. A fixed learning rate of 0.0001, Adam optimizer, and cross-entropy loss were used to train DPSeq on NCT-CRC-HE-100K. To learn histologic information without overfitting to the tissue classification problem, the model was fine-tuned for two epochs.

After fine-tuning the DPSeq, the last classifier layer was replaced with new linear layers to enable binary classification of CRC biomarkers. The biomarker classifiers were trained for up to 50 epochs with early stopping and a patience of 8, using a cosine annealing learning rate initially set to 0.0001. Weighted
cross-entropy loss was employed to handle class-imbalanced data. Finally, patient-level biomarker score was obtained by averaging the tile-level scores of all tiles in the corresponding whole-slide images.

Computational Complexity

DPSeq’s core elements consist of sequencer blocks, which replace the multihead attention in the transformer blocks of ViT. These sequencer blocks incorporate a BiLSTM2D, along with a LayerNorm layer, a BiLSTM2D module, another LayerNorm layer, and a channel-level multilayer perceptron layer, following a specific order.

The BiLSTM2D module is composed of two BiLSTM modules: BiLSTMver and BiLSTMhor. Assuming the input of BiLSTM2D as \( x \in \mathbb{R}^{H \times W \times C} \), where \( H \), \( W \), and \( C \) represent the height, width, and channel dimensions, respectively. BiLSTMver is computed \( W \) times, whereas BiLSTMhor is computed \( H \) times. The input dimension of BiLSTMver is \( H \times C \), and the hidden feature dimension \( D \) of BiLSTMver is \( \alpha C \), where \( \alpha \) is a constant.

As a result, the computational complexity of BiLSTMver is \( O(HC^2) \), and that of BiLSTMhor is \( O(WC^2) \). The combined computational complexity of these two BiLSTM modules is \( O(HC^2) \times O(W) + O(WC^2) \times O(H) = O(HWC^2) \). The computational complexity of channel fusion at the last layer of BiLSTM2D and the channel multilayer perceptron at the end of the sequencer block is also \( O(HWC^2) \). Consequently, the overall computational complexity of BiLSTM2D is \( O(HWC^2) \).

The transformer blocks used in ViT and Swin-T are multihead self-attention module and Windows multihead self-attention module, respectively. The computational complexity for multihead self-attention and Windows multihead self-attention is \( O(2HW^2 + (HW)^2\alpha C) \) and \( O(2HW^2 + M^2HWC) \), respectively, where \( M \) is the size of the small patches and exhibits resemblances to DPSeq’s sequencer block. More important, as the values of \( H \) and \( W \) increase, the computational complexity of multihead self-attention grows more rapidly compared with that of BiLSTM2D in DPSeq.

For CNN models, the computational complexity of the ResNet block in ResNet18 and ResNet50 is \( O(HWCDk^2) \), where the kernel size is \( k \times k \), the input channel size is \( C \), and the output channel size is \( D \). However, MobileNetV2 and EfficientNet showcase superior computational efficiency compared with ResNet. MobileNetV2 achieves this by introducing inverted residuals in its structure, whereas EfficientNet employs the deep scalable composite scaling method. Generally, the CNN models are less computationally complex compared with DPSeq and transformer models.

Regarding memory cost, for a single standard BiLSTM, there are cell states of size \( LC/2 \), where \( C \) is the dimension of input features, and \( L \) is the length of tokens fed into the BiLSTM. Therefore, a BiLSTM2D with one vertical and one horizontal BiLSTM used cell states of dimensions \((WC + HC)/2\), where \( W \) and \( H \) represent the width and height of the input image, respectively.

In comparison, the multihead attention in ViT involves an attention map with dimensions \( h \times (HW)^2 \), where \( h \) represents the number of attention heads. For higher resolution images, the BiLSTM2D module saved more memory compared with ViT’s multihead attention.

Experiments to compare model complexity and time efficiency were also performed.

Experiment Design

To assess the performance of the DPSeq classifier, three experiments were devised: i) comparison with published models using internal cross-validation based on TCGA-CRC-DX data set; ii) comparison with published models via external, cross-cohort validation using the same testing dataset (TCGA-CRC-DX); and iii) comparison of DPSeq with other backbone networks using the same training and testing data sets.

Comparison Using Internal Validation

Deep-learning models have been developed using internal validation to predict key CRC biomarkers (including MSI, CIMP, hypermutation, CING, BRAF mutation, and TP53 mutation) based on the whole-slide images from TCGA-CRC-DX. The study adopted the same fourfold cross-validation in the literature and compared the DPSeq’s prediction performance with the published state-of-the-art results. The data set was split into four folds in the same way as Kather et al. and/or Bilal et al. For each training iteration of the fourfold cross-validation, three folds for training and validation data sets were randomly separated with a ratio of 0.85:0.15, and the remaining fold was used for testing.

Comparison Using Cross-Cohort External Validation

To assess the generalization and robustness of DPSeq, an external validation using cross-cohort analysis was conducted. DPSeq was trained using the MCO-CRC data set (\( n = 1138, 1026, \) and 364) to predict MSI, BRAF mutation, and CIMP, respectively, and then tested the model on the TCGA-CRC-DX data set (\( n = 425, 500, \) and 235) that was unseen during the training. The testing subset of TCGA-CRC-DX has been commonly used as external testing data set in literature and can provide relatively fair comparison of performance among different published models through the external cross-cohort validation. DPSeq was compared with seven published models, including Swin-T [trained on MCO-CRC (\( n = 1065 \) for MSI status, and \( n = 1026 \) for BRAF mutation)], EfficientNet [trained on the Darmkrebs: Chancen der Verhütung durch Screening dataset (DACHS) (\( n = 2069 \) for MSI status and BRAF mutation)], ViT [trained on DACHS (\( n = 2069 \) for MSI status and BRAF mutation)].
Comparison of DPSeq with CNN Networks and Transformers

To avoid potential bias in comparing models and networks due to different training sets used in published articles, an external cross-cohort validation was conducted. This involved the use of TCGA-CRC-DX data set and other external data sets such as ResNet18 for MSI status, BRAF mutation, TP53 mutation, and CIMP status, and ShuffleNet for MSI status, BRAF mutation, and CIMP status.

Patient-level AUROC and AUPRC with ± SD of fourfold cross-validation of six key biomarkers or categories (MSI status, hypermutation, CIMP status, BRAF mutation, TP53 mutation, and CING) by DPSeq and other published models. Best results are in bold.

AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; CIMP, CpG island methylator phenotype; CING, chromosomal instability; CRC, colorectal cancer; MSI, microsatellite instability; NaN, missing value; TCGA, The Cancer Genome Atlas.
conducted using the same training and testing data sets. Specifically, four popular CNN models (ResNet18, ResNet50, MobileNetV2, and EfficientNet) and two transformers (ViT and Swin-T) were trained for prediction of MSI, CIMP, and BRAF mutation using the MCO-CRC data set and evaluated their external predictive performance using the same TCGA-CRC-DX data set. In addition to comparing the model predictive performance, the model efficiency was analyzed by measuring training time per epoch and prediction time for MSI status across all patients in TCGA-CRC-DX.

### Statistical Analysis

To assess the predictive performance of the models, the study computed the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). For the fourfold cross-validation experiment, the study obtained the average AUROC and AUPRC values across the four test folds and calculated their SD. For the external validation experiment, the study computed the AUROC and AUPRC values for the TCGA-CRC-DX data set and estimated their 95% CIs using the bootstrap method (1000 iterations).

### Results

Comparison with Published Models Using Fourfold Cross-Validation

DPSeq was used to predict six clinically relevant biomarkers for CRC (namely, MSI, CIMP, hypermutation, BRAF mutation, TP53 mutation, and CING) through fourfold cross-validation on TCGA-CRC-DX data set. To facilitate the comparison of the performance of DPSeq with previous published models, the study adhered to the same TCGA-CRC-DX split as the earlier publications. Table 1 presents the mean AUROC/AUPRC values from the fourfold cross-validation along with their corresponding SDs. DPSeq achieved the highest AUROC values of 92% (±3%) and AUPRC values of 68% (±12%) for MSI status prediction, surpassing the results reported in three recent publications.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data set for training</th>
<th>AUROC (MSI status), %</th>
<th>AUROC (BRAF mutation), %</th>
<th>AUROC (CIMP status), %</th>
<th>AUPRC (MSI status), %</th>
<th>AUPRC (BRAF mutation), %</th>
<th>AUPRC (CIMP status), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShuffleNet20</td>
<td>QUASAR (N = 1016, NaN, NaN)</td>
<td>76 (70–79)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ShuffleNet20</td>
<td>DACHS (N = 2013, NaN, NaN)</td>
<td>77 (73–79)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ShuffleNet20</td>
<td>NLCS (N = 2197, NaN, NaN)</td>
<td>72 (71–78)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>EfficientNet6</td>
<td>DACHS (N = 2069, 2069, NaN)</td>
<td>88 (83–93)</td>
<td>81 (75–86)</td>
<td>NaN</td>
<td>54 (44–63)</td>
<td>36 (25–49)</td>
<td>NaN</td>
</tr>
<tr>
<td>ViT6</td>
<td>DACHS (N = 2069, 2069, NaN)</td>
<td>81 (84–93)</td>
<td>79 (72–84)</td>
<td>NaN</td>
<td>67 (56–7)</td>
<td>30 (22–41)</td>
<td>NaN</td>
</tr>
<tr>
<td>ResNet1819</td>
<td>Pooled data sets (N = 7917)</td>
<td>91 (87–95)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>Swin-T7</td>
<td>MCO (N = 1065, 1026, NaN)</td>
<td>90 (85–95)</td>
<td>80 (73–86)</td>
<td>76 (68–84)</td>
<td>72 (61–82)</td>
<td>39 (28–54)</td>
<td>NaN</td>
</tr>
<tr>
<td>DPSeq (ours)</td>
<td>MCO (N = 1138, 1026, 364)</td>
<td>89 (83–94)</td>
<td>83 (77–88)</td>
<td>80 (72–87)</td>
<td>71 (59–83)</td>
<td>46 (34–61)</td>
<td>63 (51–75)</td>
</tr>
</tbody>
</table>

Patient-level AUROC and AUPRC with a 95% CI obtained via bootstrapping (×1000) calculated DPSeq and other published models in predicting MSI, BRAF mutation, and CIMP. Best results are in bold. The second column contains the data set for training, and their number of training samples for three biomarkers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data set for training</th>
<th>AUROC (MSI status), %</th>
<th>AUROC (BRAF mutation), %</th>
<th>AUROC (CIMP status), %</th>
<th>AUPRC (MSI status), %</th>
<th>AUPRC (BRAF mutation), %</th>
<th>AUPRC (CIMP status), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShuffleNet20</td>
<td>QUASAR (N = 1016, NaN, NaN)</td>
<td>76 (70–79)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ShuffleNet20</td>
<td>DACHS (N = 2013, NaN, NaN)</td>
<td>77 (73–79)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ShuffleNet20</td>
<td>NLCS (N = 2197, NaN, NaN)</td>
<td>72 (71–78)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>EfficientNet6</td>
<td>DACHS (N = 2069, 2069, NaN)</td>
<td>88 (83–93)</td>
<td>81 (75–86)</td>
<td>NaN</td>
<td>54 (44–63)</td>
<td>36 (25–49)</td>
<td>NaN</td>
</tr>
<tr>
<td>ViT6</td>
<td>DACHS (N = 2069, 2069, NaN)</td>
<td>81 (84–93)</td>
<td>79 (72–84)</td>
<td>NaN</td>
<td>67 (56–7)</td>
<td>30 (22–41)</td>
<td>NaN</td>
</tr>
<tr>
<td>ResNet1819</td>
<td>Pooled data sets (N = 7917)</td>
<td>91 (87–95)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>Swin-T7</td>
<td>MCO (N = 1065, 1026, NaN)</td>
<td>90 (85–95)</td>
<td>80 (73–86)</td>
<td>76 (68–84)</td>
<td>72 (61–82)</td>
<td>39 (28–54)</td>
<td>NaN</td>
</tr>
<tr>
<td>DPSeq (ours)</td>
<td>MCO (N = 1138, 1026, 364)</td>
<td>89 (83–94)</td>
<td>83 (77–88)</td>
<td>80 (72–87)</td>
<td>71 (59–83)</td>
<td>46 (34–61)</td>
<td>63 (51–75)</td>
</tr>
</tbody>
</table>
Moreover, DPSeq exhibited significant improvements over the published models in predicting hypermutation and CIMP status. For hypermutation prediction, DPSeq’s average AUROC and AUPRC values of 88% (±3%) and 65% (±5%), respectively, were 3% to 8% higher than the results reported by Guo et al. 7, Bilal et al. 17, and Kather et al. 18 In CIMP status prediction, DPSeq achieved AUROC and AUPRC values of 81% (±4%) and 65% (±4%), respectively, which were 4% to 14% higher than the results reported in previous publications. Furthermore, DPSeq’s predictive performance for BRAF mutation, TP53 mutation, and CING status was also competitive with state-of-the-art results. Specifically, DPSeq achieved an AUPRC value of 38% (±3%) for BRAF mutation prediction, which was 3% to 5% higher than the results reported by Guo et al 7 and Bilal et al. 17

Comparison with Published Models Using Cross-Cohort External Validation

DPSeq was trained using the MCO-CRC data set to predict MSI status, BRAF mutation, and CIMP status. The robustness of DPSeq’s predictive performance was tested using TCGA-CRC-DX data sets for these three biomarkers. DPSeq demonstrated exceptional performance in external validation on TCGA-CRC-DX data sets (Figure 2). DPSeq’s performance in predicting MSI status resulted in an AUROC value of 89% (95% CI, 83%—94%) and an AUPRC value of 71% (95% CI, 60%—83%). When predicting BRAF mutation on TCGA-CRC, DPSeq achieved an AUROC value of 83% (95% CI, 77%—88%) and an AUPRC value of 46% (95% CI, 34%—61%). Regarding CIMP status, DPSeq demonstrated an AUROC value of 80% (95% CI, 72%—87%) and an AUPRC value of 63% (95% CI, 51%—75%).

The same TCGA-CRC-DX data sets were used in the published articles for external validation of these biomarkers, 6, 7, 19, 20 which allows fair comparison of DPSeq’s predictive performance with published models (Table 2). However, more training data can usually improve the predictive performance in external validation data sets. For predicting BRAF mutation and CIMP status, DPSeq significantly outperforms all other methods, even when trained with small subsets of MCO-CRC (n = 1026 for BRAF mutation, and n = 364 for CIMP status). DPSeq exhibited superior performance in predicting MSI status compared with ViT and EfficientNet and achieved similar results to current state-of-the-art models, including the recent Swin-T model 7 and the Resnet18 model developed by Echle et al. 19 The Resnet18 model was trained on a significantly larger, multicenter data set (N = 7917), which is almost eight times larger than the MCO-CRC data set (N = 1138).

Comparison with CNN and Transformer Models

To ensure a fair comparison, four popular CNN models (ResNet18, ResNet50, MobileNetV2, and EfficientNet) and two advanced transformer models (Vision Transformer (ViT) and Swin Transformer (Swin-T)) were trained in the same way as DPSeq using the MCO-CRC data set. The external predictive performance of
these models was compared using the TCGA-CRC-DX data set. As depicted in Figure 3, DPSeq surpassed all other CNN and transformer models, achieving the highest AUROC and AUPRC values in all prediction tasks (MSI status, \textit{BRAF} mutation, and CIMP status). In predicting MSI status, DPSeq achieved an AUROC value of 89% (95% CI, 83%–94%) and an AUPRC value of 71% (95% CI, 60%–83%), which was about 3% higher than the transformer models and >10% to 28% higher than the CNN models. Similarly, DPSeq’s AUROC and AUPRC values in predicting \textit{BRAF} mutation [AUROC = 83% (95% CI, 77%–88%) and AUPRC = 46% (95% CI, 34%–61%)]
were at least 5% and 2% higher than ViT and Swin-T, respectively. Notably, DPSeq’s AUPRC value was about 9% higher than that of ViT (AUPRC = 37% (95% CI, 26.6%–52.4%)). In predicting CIMP status, DPSeq slightly outperformed Swin-T but significantly outperformed other models in terms of the AUROC results. Moreover, DPSeq’s AUPRC value [63% (95% CI, 51.6%–75.0%)] for predicting CIMP was approximately 8% and >20% higher than ViT and the CNN models, respectively.

Model Complexity and Time Efficiency

DPSeq was compared with reference networks not only in terms of predictive performance, but also in terms of model complexity and time efficiency. As expected, larger models generally required longer training and prediction times (Figure 4). Despite being larger than CNN models, DPSeq was smaller than transformers and required less training and prediction time. Moreover, DPSeq achieved superior predictive performance compared with transformer models. In contrast, although CNN-based models were much smaller and faster, their predictive performance was significantly inferior to that of DPSeq. Therefore, taking into account all three factors of model complexity, time efficiency, and predictive performance, DPSeq appears to be the preferred choice for practical applications.

Discussion

CNNs (eg, ResNet, MobileNetV2, and EfficientNet) have become the dominant architecture in digital pathology, for tasks such as tumor detection, subtyping, and grading, and predicting molecular biomarkers using H&E-stained histopathologic images. More recently, vision transformers have emerged and surpassed CNNs. However, transformers are usually extremely complex and resource-demanding because of their large model size and number of parameters.

This study developed a novel and efficient digital pathology classifier called DPSeq to predict cancer biomarkers through fine-tuning a sequencer architecture integrating horizontal and vertical BiLSTM networks. Several articles have adopted BiLSTM models in the fields of computer vision and digital pathology as well. Dubey et al devised an innovative integrated architecture that merges a CNN based on the Inception-V3 model with BiLSTM models, incorporating a self-attention mechanism. Similarly, Tripathi et al employed BiLSTM models to leverage contextual information within patches originating from the same region for tumor classification. Aslan developed a hybrid structure that combines both CNN and BiLSTM for mammography image classification. In addition, Yao et al used a parallel structure, leveraging a CNN alongside a BiLSTM for image feature extraction. Unlike previous research, DPSeq adopted a unique approach to using BiLSTM. It incorporated a newly introduced BiLSTM2D module, which included both vertical and horizontal BiLSTM layers, within the sequencer block. This BiLSTM2D module replaced the traditional multihead self-attention module commonly found in transformer blocks. The innovative design of the sequencer block enabled the BiLSTM to be used multiple times within a deep neural network, potentially resulting in better performance.

On the basis of H&E-stained histopathologic images, DPSeq demonstrated exceptional performance for predicting key biomarkers in CRC (MSI status, hypermutation, CIMP status, BRAF mutation, TP53 mutation, and CING), outperforming most published state-of-the-art models in a within-cohort internal validation and a cross-cohort external validation. In addition, under the same experimental conditions using the same set of training and testing data sets, DPSeq surpassed four CNN (ResNet18, ResNet50, MobileNetV2, and EfficientNet) and two transformer (ViT and Swin-T) models, achieving the highest AUROC and AUPRC values in predicting MSI status, BRAF mutation, and CIMP status. Compared with current CNN and transformer models, DPSeq reduced the model size to a level comparable to that of ResNet50, while providing better or similar prediction performance than the larger and more complex transformer models, such as ViT and Swin-T. Our experiments showed that DPSeq required less time for training and prediction than transformer models. Overall, DPSeq demonstrated the highest performance/complexity ratio among all CNN- and transformer-based models tested. Therefore, DPSeq appears to be the preferred choice over transformer and CNN models for predicting cancer biomarkers. The advantages of BiLSTM networks indicate that they could be a promising and practical backbone for digital pathology tasks. As such, additional research and innovation on BiLSTM architectures should be pursued in the areas of computer vision and digital pathology.

Pathology images, particularly WSIs, pose a considerable challenge because of their immense size, containing billions of pixels. Therefore, WSIs are often divided into smaller image patches for training neural networks to predict specific outcomes. However, given the lack of patch-level labels, constructing a direct prediction model at the patch level is infeasible. Currently, two popular strategies have been developed: multiple-instance learning and the classic weakly supervised approach. Laleh et al have demonstrated that the classic weakly supervised approach was employed across all the neural networks examined in this article to ensure fair comparisons.
Finally, a key distinction in DPSeq’s approach lies in its communication among local regions compared with ViT. Instead of establishing communication between all local pairs, DPSeq employed horizontal and vertical BiLSTM to consider the relationship between pixels in a cross-shaped expansion. This BiLSTM extracted advanced horizontal and vertical features. Through feature combination and channel fusion, DPSeq further learned the relationship between rows and columns, ensuring that long-distance relationships across the entire graph were taken into account. This approach proves to be more efficient than considering pairwise tokens’ relationships, making it well suited for tissue texture images, such as histopathologic images.

Disclosure Statement
X.S.X. is an employee of Genmab, Inc. Genmab did not provide any funding for this study.

Supplemental Data
Supplemental material for this article can be found at http://doi.org/10.1016/j.ajpath.2023.09.006.

References