Computerized Medical Imaging and Graphics 87 (2021) 101832

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Machine learning techniques for mitoses classification

Shima Nofallah [a], Sachin Mehta [a], Ezgi Mercan [a], Stevan Knezevich [b], Caitlin J. May [a], Donald Weaver [c], Daniela Witten [a], Joann G. Elmore [d,1], Linda Shapiro [a,1,*]

[a] University of Washington, Seattle WA 98195, USA
[b] Pathology Associates, Clovis, CA 983611, USA
[c] University of Vermont, Burlington VT 05405, USA
[d] David Geffen School of Medicine, UCLA, Los Angeles CA 90024, USA

ARTICLE INFO

Keywords:
Pathology
Mitoses
Melanoma
Convolutional neural networks
Machine learning

ABSTRACT

Background: Pathologists analyze biopsy material at both the cellular and structural level to determine diagnosis and cancer stage. Mitotic figures are surrogate biomarkers of cellular proliferation that can provide prognostic information; thus, their precise detection is an important factor for clinical care. Convolutional Neural Networks (CNNs) have shown remarkable performance on several recognition tasks. Utilizing CNNs for mitosis classification may aid pathologists to improve the detection accuracy.

Methods: We studied two state-of-the-art CNN-based models, ESPNet and DenseNet, for mitosis classification on six whole slide images of skin biopsies and compared their quantitative performance in terms of sensitivity, specificity, and F-score. We used raw RGB images of mitosis and non-mitosis samples with their corresponding labels as training input. In order to compare with other work, we studied the performance of these classifiers and two other architectures, ResNet and ShuffleNet, on the publicly available MITOS breast biopsy dataset and compared the performance of all four in terms of precision, recall, and F-score (which are standard for this data set), architecture, training time and inference time.

Results: The ESPNet and DenseNet results on our primary melanoma dataset had a sensitivity of 0.976 and 0.968, and a specificity of 0.987 and 0.995, respectively, with F-scores of .968 and .976, respectively. On the MITOS dataset, ESPNet and DenseNet showed a sensitivity of 0.866 and 0.916, and a specificity of 0.973 and 0.980, respectively. The MITOS results using DenseNet had a precision of 0.939, recall of 0.916, and F-score of 0.927. The best published result on MITOS (Saha et al. 2018) reported precision of 0.92, recall of 0.88, and F-score of 0.90. In our architecture comparisons on MITOS, we found that DenseNet beats the others in terms of F-Score (DenseNet 0.927, ESPNet 0.890, ResNet 0.865, ShuffleNet 0.847) and especially Recall (DenseNet 0.916, ESPNet 0.866, ResNet 0.807, ShuffleNet 0.753), while ResNet and ESPNet have much faster inference times (ResNet 6 s, ESPNet 8 s, DenseNet 31 s). ResNet is faster than ESPNet, but ESPNet has a higher F-Score and Recall than ResNet, making it a good compromise solution.

Conclusion: We studied several state-of-the-art CNNs for detecting mitotic figures in whole slide biopsy images. We evaluated two CNNs on a melanoma cancer dataset and then compared four CNNs on a public breast cancer data set, using the same methodology on both. Our methodology and architecture for mitosis finding in both melanoma and breast cancer whole slide images has been thoroughly tested and is likely to be useful for finding mitoses in any whole slide biopsy images.

1. Introduction

Melanomas account for approximately 75 % of all skin-cancer-related deaths and are responsible for over 10,000 deaths annually in the United States alone (Esteva et al., 2017). Melanoma is highly curable when detected in its earliest stage (Society, 2016). The gold standard for

* Corresponding author at: 634 Paul G, Allen Center for Computer Science & Engineering, 185 E Stevens Way NE, Seattle, WA 98195, USA.
E-mail addresses: shima@cs.washington.edu (S. Nofallah), sacmehta@cs.washington.edu (S. Mehta), ezgi@cs.washington.edu (E. Mercan), shapiro@cs.washington.edu (S. Knezevich), caitmay@u.washington.edu (C.J. May), donald.weaver@uvmhealth.org (D. Weaver), dwitten@uw.edu (D. Witten), jelmore@mednet.ucla.edu (J.G. Elmore), shapiro@cs.washington.edu (L. Shapiro).
[1] These authors share senior authorship.

https://doi.org/10.1016/j.compmedimag.2020.101832
Received 1 January 2020; Received in revised form 9 October 2020; Accepted 17 November 2020
Available online 27 November 2020

diagnosis of melanoma is the histopathological examination in which the skin biopsy specimen is examined under a microscope by a pathologist (Cireşan et al., 2013). However, a single whole slide image of one tissue sample has a size of approximately 2.2 Gigapixels and the biopsy material often includes more than one tissue section with hundreds of thousands of cells on each slide, posing a great challenge for the pathologist to fully analyze all of the cellular data within the images. A pathologist's diagnosis is often subjective and prone to variability (Elmore et al., 2015; Elmore et al., 2017); automated diagnosis holds promise to improve accuracy and reproducibility (Mercan et al., 2019). Thus, research on the automated classification of skin biopsies has gained traction with the overall goal of assisting pathologists to make accurate diagnoses.

Melanoma diagnosis involves histological analysis of various cellular and architectural features. Melanocytic lesions range across a broad spectrum of categories: 1) benign, 2) variably atypical (e.g. demonstrating mild, moderate or severe atypia), 3) melanoma in situ, 4) invasive melanoma stage T1a, and 5) invasive melanoma >= stage T1b (Piepkorn et al., 2014). A mitosis (or mitotic figure) remains an important entity in the review of skin biopsy cases as their presence may aid in the diagnosis of a melanoma in addition to being associated with poorer prognosis. A high mitotic rate in a primary invasive melanoma is associated with a lower survival probability. Among the independent predictors of melanoma-specific survival, mitotic rate is the strongest prognostic factor after tumor thickness (Thompson et al., 2011). Thus, the accurate detection of mitotic activity is an important role for the pathologist in making cancer diagnoses, and because mitoses are small objects with various shapes that can resemble normal nuclei, mitosis detection remains a challenging task for humans. Because of its clinical importance, the development of automated mitosis detection has become an active area of research with the goal of developing decision support systems to assist pathologists (Li et al., 2018a).

Various approaches have been applied to detect mitotic figures. Sertel et al. (2009) computed the probability map based on the likelihood functions and then used a component-wise two-step thresholding to find mitoses in neuroblastoma. A graph-based multi-resolution approach with color and texture features was used by Roullier et al. (2010), Roux et al. (2013) for mitosis extraction in breast biopsy images. Irshad et al. used morphological features to identify cellular entities in a breast biopsy dataset (Irshad et al., 2013).

In recent years, with the development of fast and accessible Graphics Processing Units (GPUs), Convolutional Neural Networks (CNNs) have gained attention for medical image analysis, primarily because of their capability to learn strong structural representations about objects of interest (e.g. cellular entities (Cireşan et al., 2013) or tissues (Mehta et al., 2018a; Ronneberger et al., 2015)). For example, Cireşan et al. (2013) used a CNN-based method for mitosis detection and won the International Conference on Pattern Recognition 2012 (ICPR 2012) mitosis detection challenge by a significant margin. Since then, much of the research on mitosis detection in breast cancer biopsy images has used CNNs. Simo-Serra et al. (2015), Irshad et al. (2013) and Wang et al. (2014) developed different methods that merge CNN image descriptors and handcrafted features to improve the detection. Chen et al. (2016) proposed a two-stage mitosis detection pipeline, with a coarse retrieval model, followed by a fine discrimination model. In recent work, Li et al. (2018b) used a deep detection network using residual connection when only the weak label is given. López-Tapia et al. (2019) introduced a pyramidal model to detect mitoses. On each pyramid level, a Bayesian convolutional neural network is trained to compute class prediction and uncertainty on each pixel.

Several CNN-based methods have been proposed for mitosis detection in different tissues, including breast (Cireşan et al., 2013; Irshad et al., 2013; Chen et al., 2016), stem cells (Zhou et al., 2017), and skin (López-Tapia et al., 2019). Unlike natural image datasets (e.g. the ImageNet Deng et al., 2009), the number of training samples are limited in medical image datasets usually by an order of a few hundred (Roullier et al., 2010; Veta, 2016; Veta et al., 2015). To achieve strong performance on these datasets, CNNs have been complemented with several methods, including hand-crafted features (Saha et al., 2018; Irshad et al., 2013; Dodballapur et al., 2019) and better augmentation strategies (Ronneberger et al., 2015). U-Net (Ronneberger et al., 2015) introduced an encoder-decoder architecture with skip-connections for segmenting different biological structures in images and demonstrated good performance across several datasets

Most research in mitosis detection has been conducted on biopsy images other than the skin (Saha et al., 2018; Mercan et al., 2019; Mehta et al., 2018a; Chen et al., 2016). However, skin biopsy images are different from these biopsy images in terms of texture, color, and mitosis shape, as shown in Fig. 1. As a result, existing CNN-based classifiers trained on these biopsy images may have poor performance on skin biopsies. Moreover, to the best of our knowledge, there are no publicly available skin biopsy datasets with mitosis annotations. Given the importance of mitosis detection in skin cancer diagnosis, we created a new dataset with mitosis-level markings from an expert pathologist. We studied and compared the performance of two different state-of-the-art CNNs, one that is lightweight in terms of parameters and execution time and one that is much bigger, in terms of accuracy, sensitivity, specificity, precision, recall, and F-score. We then compare the performance of these two CNNs with two additional state-of-the-art architectures on a public breast cancer data set in terms of precision, recall, F-score, architecture, training time, and inference time. Our work has several contributions: 1) This is the first paper to experiment with finding mitotic figures in whole slide melanoma biopsies. 2) After determining the best possible performances on the melanoma biopsy slide images, we showed that this pipeline could be applied to a well-known breast cancer data set (MITOS) and compared the results from our two models (ESPNet, which was chosen for lightweight network and speed, and DenseNet, which was an example of a state-of-the-art network) with the results from several published papers, showing that DenseNet could beat all of them and ESPNet came close (Table 4). 3) We ran two more models, ResNet and ShuffleNet, on the MITOS dataset for further comparison and found that DenseNet is still the best performer in terms of F-1 score (DenseNet 0.927, ESPNet 0.890, ResNet 0.865 and ShuffleNet 0.847) and, particularly, in terms of Recall (DenseNet 0.916, ESPNet 0.866, ResNet 0.870 and ShuffleNet 0.753), which is very important for cancer grading. 4) Our paper, in general, gives a methodology and architecture for mitosis finding in both melanoma and breast cancer whole slide images, and that is likely to be useful for finding mitoses in any whole slide biopsy images.

2. Materials and methods

2.1. Dataset and preprocessing

Our dataset comes from hematoxylin and eosin (H&E) stained slides of skin biopsy images, acquired in the MPATH study (R01 CA151306). The Institutional Review Board at the University of Washington approved all test set study activities. The identification and development of these images has been previously described in (Elmore et al., 2015). All glass slides of skin biopsies were scanned at 40x magnification with a high-quality digital scanner. The compression method we used on these images is tiff.

2.1.1. Dataset and materials

An experienced pathologist (SK) chose six skin biopsy cases of >= pT1b invasive melanoma, a diagnostic category known to be associated with high mitotic activity, from our dataset and cropped 34 areas in the whole slide images (WSIs) of these cases. The size of the areas and the number of areas per each case were not fixed but were based on the pathologist's judgment with the aim of marking as many mitoses as possible. A total of 628 mitoses in the cropped image areas were marked by the same pathologist with a green dot on each mitosis, using the

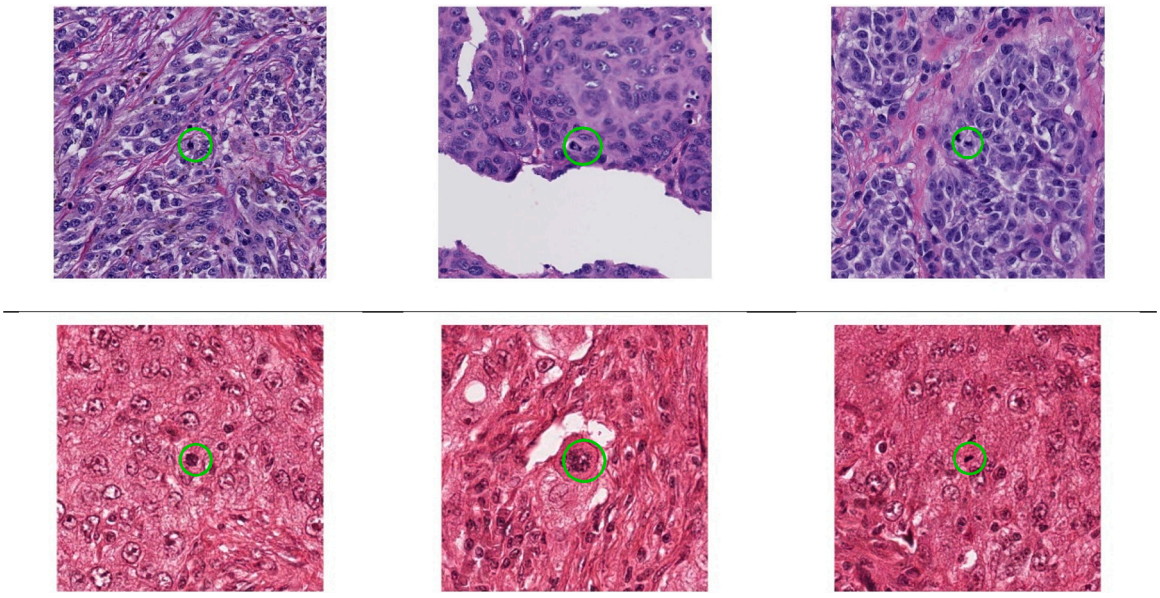

Fig. 1. Example crops of biopsy images with mitoses in them; **(top)** skin; **(bottom)** breast. These biopsies are different in terms of color, texture, and mitosis phase and shape.
*A mitosis in each image is present near the center and is marked with a green circle for visualization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Sedeen Viewer (Martel et al., 2017). These marked mitoses provide "class mitosis" samples for training and validation of our binary classifiers. The details about our skin biopsy dataset are summarized in Table 1.

Distinguishing mitoses from normal nuclei is a challenge for automated mitosis classifiers. Mitoses and nuclei can appear very similar in color and shape; thus, the classifiers require a large number of nuclei samples to differentiate between these cellular entities. If the whole non-mitosis regions of the image were to be sampled uniformly, many of the non-interesting instances such as background would be in the class "non-mitosis" and training a strong classifier would be inefficient. To avoid this, we used a standard watershed-based nuclei segmentation method (Corredor et al., 2018) to find nuclei in the images and use them as examples for the class non-mitosis. Fig. 2 shows the output of this nuclei detector on a cropped portion of a skin biopsy.

Fig. 3 shows some examples of mitoses and normal nuclei, which we note are very similar in terms of texture, color, and shape. In the process of sampling mitoses and nuclei, based on our experiments, we used a 101×101 patch approximately centered on the target entity's center. If a part of this window lies outside of the image borders, the image is padded using mirroring of the border pixels. To help our classifier learn rotation, scale, and translation-invariant representations, we augmented our training set with standard augmentation methods such as rotation (45, 90, 135 or 225 degrees) and mirroring (horizontal and vertical)

The number of mitoses per slide is an order of magnitude fewer than other entities, such as nuclei and melanocytes present in the slide. In other words, the dataset is imbalanced. If we train a classifier with such an imbalanced dataset, then the classifier will be biased towards the entities with more samples. To address this imbalance, a standard approach (Prati et al., 2009; Ren et al., 2015) is to maintain a good ratio between positive samples (patches that contain mitoses) and negative samples (patches that do not contain mitoses). For our dataset, we empirically found that this ratio is 1:3 i.e. the number of negative samples available for training is approximately 3 times the number of positive samples; resulting in 4364 mitoses and 12,640 non-mitosis samples after data augmentation. Since we used a watershed-based nuclei segmentation (Corredor et al., 2018) as a pre-processing method, non-mitosis samples mostly contain nuclei.

Table 1
Mitosis dataset summary – **Melanoma**.

Case ID	Number of slices	Number of cells in WSI	Number of areas	Number of mitoses
Case # 1	5	~ 250k	14	197
Case # 2	3	~ 237k	6	32
Case # 3	6	~ 320	7	232
Case # 4	1	~ 115k	5	156
Case # 5	3	~ 49k	1	6
Case # 6	4	~ 39k	1	5
Total		–	34	628

2.1.2. Data split

We split our dataset randomly into training (80 %) and validation (20 %) sets, respectively. The validation set was withheld during the training phase. After the training is complete, validation set is used to evaluate the trained model performance.

2.2. Training

2.2.1. Networks

Our classification network uses a standard pipeline (Krizhevsky et al., 2012; He et al., 2016) that stacks encoding and down-sampling units to learn latent representations. In our experiments, we used two state-of-the-art encoding units: 1) Efficient Spatial Pyramid of Dilated Convolutions (ESPNet) (Mehta et al., 2018b) and 2) Densely Connected Convolutional Networks (DenseNet) (Huang et al., 2017). The same dataset split was used for both ESPNet and DenseNet training and validation.

Efficient spatial pyramid of dilated convolutions (ESPNet): ESPNet (Mehta et al., 2018b) is a fast and efficient CNN that was designed for semantic segmentation on mobile devices. The core building block of the ESPNet architecture is the ESP unit that decomposes a standard convolution into a point-wise convolution and a spatial pyramid of

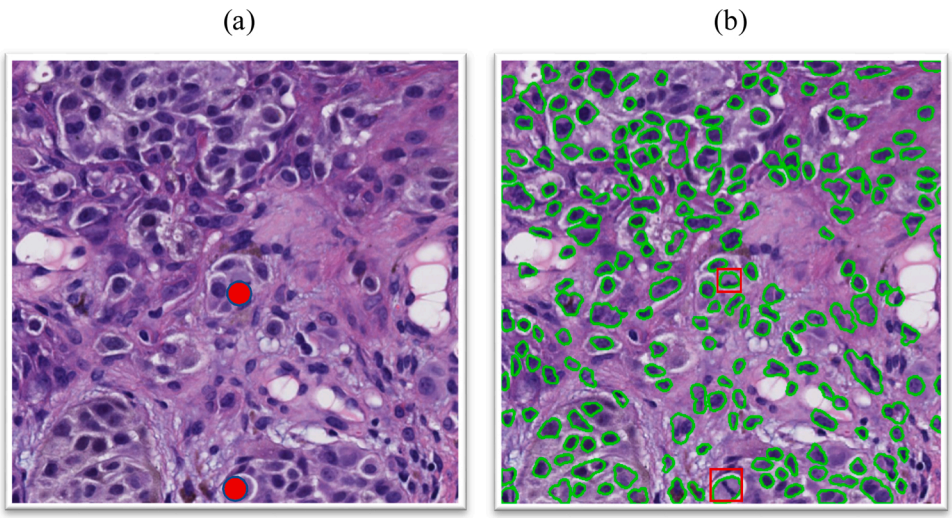

Fig. 2. Examples of applying the nuclei segmentation method (Corredor et al., 2018) on a crop of skin biopsy image **(a)** original crop **(b)** nuclei segmentation result.
* Two mitoses that are present in the original crop are marked with red dots for visualization.
* Segmentation method was able to find the mitoses. We marked them here with red boxes for visualization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

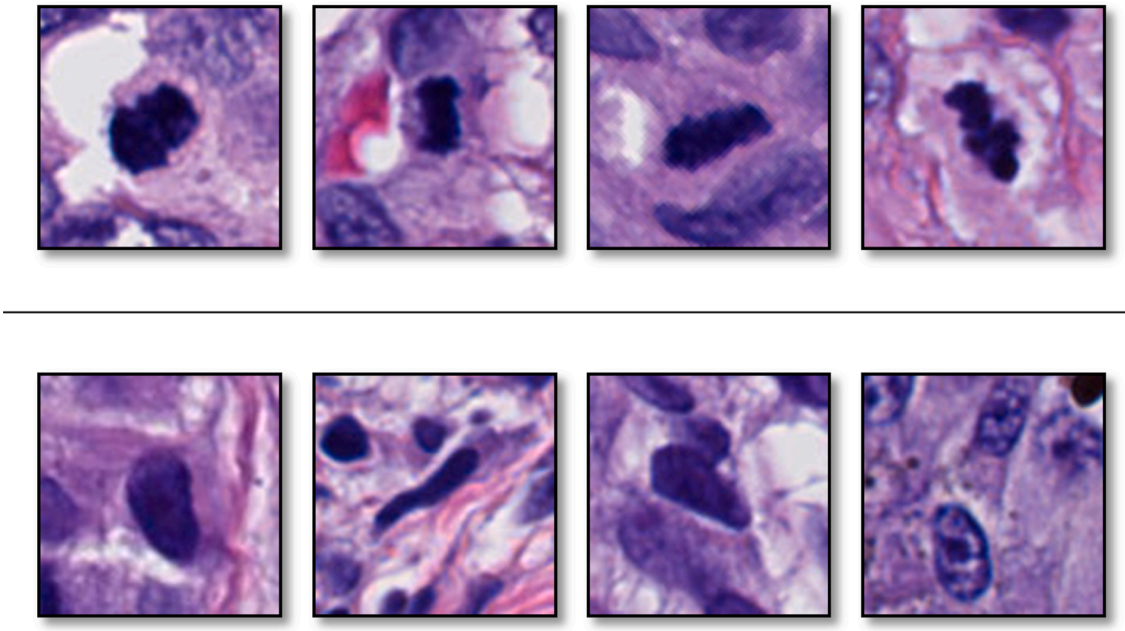

Fig. 3. Examples of **(top)** sampled mitoses, and **(bottom)** sampled nuclei that are not mitoses. These two entities have similarity in color, surrounding and texture.

dilated convolution. This factorization reduces the computational complexity of the ESP unit in comparison to the standard convolution. Fig. 4 **(a)** visualizes the ESP unit. We chose this unit in our study because of its good performance in segmenting breast biopsy whole slide images (Mehta et al., 2018a).

Densely Connected Convolutional Networks (DenseNet): DenseNet, densely connected convolutional neural network (Huang et al., 2017), introduces a novel connectivity mechanism to improve the flow of information between different stacked convolutional layers. As shown in Fig. 4 **(b)**, this unit establishes a direct link between different convolutional layers. This connectivity pattern provides multiple paths for gradients to flow back to the input and thus, helps in learning better representations.

2.2.2. Training parameters

We train our classifiers using the ADAM optimizer (Kingma and Adam, 2014) for a total of 20 epochs with an initial learning rate of 0.001. We decay the learning rate by 0.1 after every 5 epochs. During training, we minimize the cross-entropy loss (De Boer et al., 2005).

2.2.3. Evaluation metrics

We evaluate the performance of our classifier on the melanoma dataset using six metrics: four standard metrics (precision, recall, F-score, and accuracy) and two widely used metrics in clinical care (sensitivity and specificity):

- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F-score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$

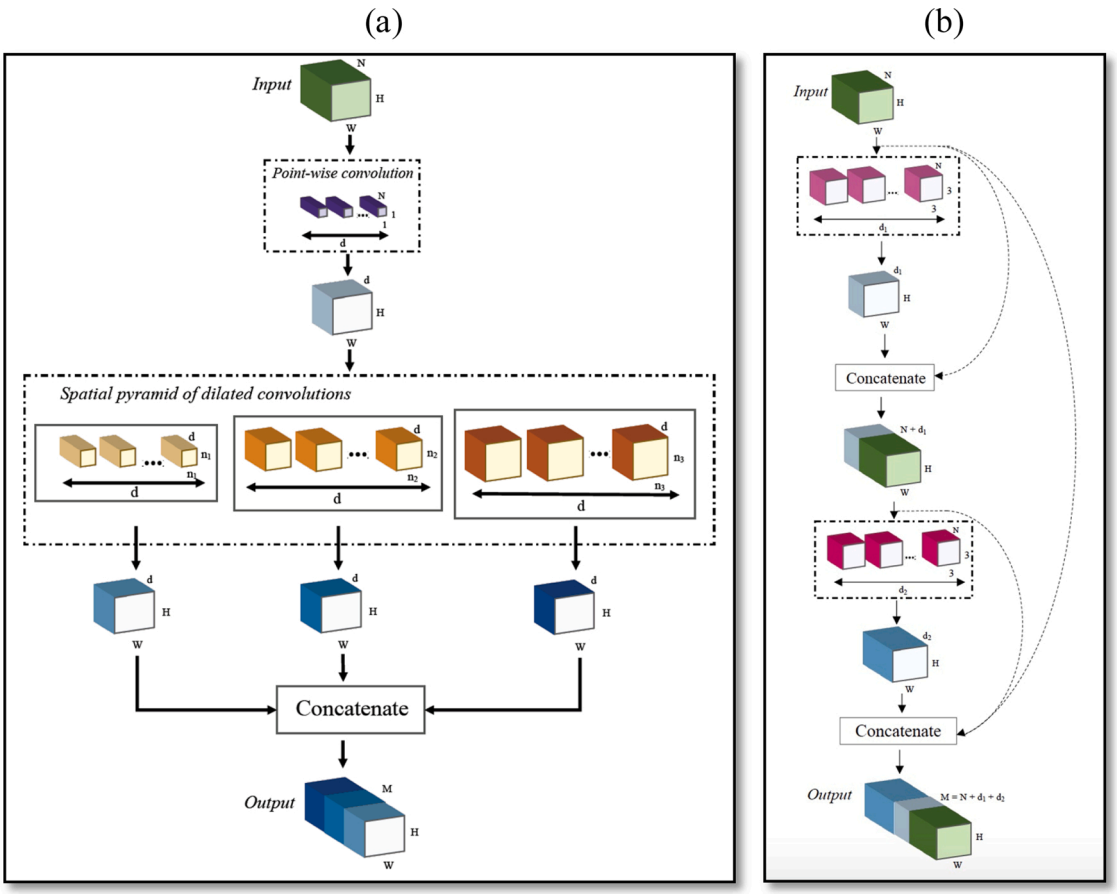

Fig. 4. Two convolutional units, ESPNet **(a)** and DenseNet **(b)**, that are used in our experiment. Each of these units receives a 3D tensor with width W, height H, and depth N as an input and produces a 3D tensor with width W, height H, and depth M as an output. The projection channel dimension in ESPNet unit is represented by d while in DenseNet unit, it is represented by d_i. For ESPNet, output tensor depth is $M = k \times d$, where k is the number of parallel branches in the ESPNet unit ($k = 3$ in **(a)**), the size of the point-wise convolution is 1×1, and n_i is the size of the dilated convolutional layers. For more information, see (Mehta et al., 2018b). For the DenseNet unit, output tensor depth is $M = \sum d_i$, $i = \{1, \ldots, L\}$, where L represents the number of stacked layers ($L = 3$ in **(b)**). It is common to use 3×3 standard convolutional layers in DenseNets. For more information, see (Huang et al., 2017).

where True Positive (*TP)* is the number of correctly predicted mitosis and True Negative (TN) is the number of correctly predicted non-mitosis samples, while False Negative (*FN)* is the number of mitosis samples which classified as non-mitosis by the classifier and False Positive (FP) are the non-mitosis samples predicted as mitosis. F-score is the harmonic mean of precision and recall.

3. Results

3.1. Mitosis detection results on Melanoma dataset

Table 2 summarizes the results of our classifiers using two different encoding units: 1) ESPNet and 2) DenseNet. Both networks achieved high accuracy on classifying mitoses with a sensitivity of 0.976 and 0.968, and specificity of 0.987 and 0.995, respectively. Though DenseNet outperformed ESPNet, this outperformance was not statistically significant (p-value is 0.5716), and the training time of ESPNet is about a third that of DenseNet (see Table 2) (Table 3).

Table 2
Quantitative results of ESPNet and DenseNet on validation set* of **Melanoma**.

Metrics	ESPNet (Mehta et al., 2018b)	DenseNet (Huang et al., 2017)
Accuracy	0.984	0.988
Precision	0.961	0.984
Recall	0.976	0.968
F-score	0.968	0.976
Sensitivity	0.976	0.968
Specificity	0.987	0.995
FP, FN	5, 3	2, 4
TP, TN	122, 370	121, 373
*Training Time***	35 minutes	106 min

* Validation set contains 20 % of the whole set (no data augmentation).
** Experiments were performed on a 2.10 GHz Intel Xeon Silver 4110 CPU with GeForce GTX 1080 GPU. Utilization of a GPU and small patch size speed up the training process. In addition, ESPNet is a much lighter model than DenseNet, which explains the lower training time of ESPNet compared to that of DenseNet. We trained our classifiers using the ADAM optimizer for a total of 20 epochs with an initial learning rate of 0.001. We decayed the learning rate by 0.1 after every 5 epochs. During the training process, we minimized the cross-entropy loss.

3.2. Generalizability of the MITOS dataset

To study the generalization ability of our classifiers on other datasets, we evaluated the performance on a publicly available mitosis dataset for breast biopsies: MITOS (Roullier et al., 2010; Roux et al., 2013). The dataset consists of 50 images corresponding to 50 high-power fields in 5 different breast cancer slides stained with hematoxylin and eosin. This dataset contains 800 mitoses.

We first compared our two classifiers (ESPNet and DenseNet) to the results reported in several papers in the recent literature (Saha et al., 2018; Cireşan et al., 2013; Li et al., 2018a; López-Tapia et al., 2019; Dodballapur et al., 2019) The architectures of these classifiers can be summarized as follows:

Table 3
Quantitative results of ESPNet and DenseNet on **MITOS** (Roullier et al., 2010).

Metrics	ESPNet	DenseNet
Accuracy	0.946	0.964
Precision	0.916	0.939
Recall	0.866	0.916
F-score	0.891	0.927
Sensitivity	0.866	0.916
Specificity	0.973	0.980
FP, FN	16, 27	12, 17
TP, TN	175, 582	185, 586

- **Saha, et al.** The deep learning consists of two parts: (1) a convolutional neural network and (2) a handcrafted feature extractor. The deep architecture contains five convolution layers, four max-pooling layers, four ReLUs, and two fully connected layers.
- **Dodballapur** et al. In this work, handcrafted features extracted from the masks generated from the Mask R-CNN network are combined with deep features to classify the candidate cells. To extract an image-level representation, the Xception network pre-trained on ImageNet without the last two fully connected layers was used.
- **Li, et al.** Their pipeline consists of three components: (1) a deep detection model (DeepDet) that produces primary detection results, (2) a deep verification model (DeepVer) that verifies these detections and eliminates false positives, and (3) a deep segmentation model (DeepSeg) that segment the images and generates bounding box annotations around segmented regions to provide weak box-level annotations. The DeepDet model consists of an RPN (Region Proposal Network) and a region-based classifier. The DeepVer model is based on the ResNet.
- **López-Tapia, et al.** Their pipeline consists of two components: first, a coarse-to-fine cascade of CNN Bayesian models for mitosis detection; then, to make the model resistant to local and shape deformations, a Spatial Transforming Layer is applied before the 4th and 7th residual blocks in scale x40.
- **Cireşan, et al.** They trained two DNNs and ensembled the performance evaluation results: DNN1 contains five convolutional layers, five max-pooling layers, and two fully connected layers. DNN2 contains four convolutional layers, four max-pooling layers, and two fully connected layers.

For comparison, the architectures of ESPNet and DenseNet are as follows:

- **ESPNet**: Our classification network uses a standard pipeline that stacks encoding and down-sampling units to learn latent representations. The model contains one conventional 2D convolution layer, five ESP blocks, four down-sampling layers, one average-pooling, and two fully connected layers.
- **DenseNet**: We used the DenseNet161 architecture which contains one conventional 2D convolution layer, four Dense block, three Transition layers, one max-pooling, and two fully connected layers.

In comparison to existing state-of-the-art methods (see Table 4), our classifiers achieve a competitive performance. In particular, our DenseNet-based classifier is 2% more accurate than Saha et al. (2018).

In order to compare more thoroughly, we added two more state-of-the-art CNNs, ResNet (He et al., 2016) and ShuffleNet (Zhang et al., 2018) to the original two (ESPNet and DenseNet). We compared all four classifiers on precision, recall, and F-score (as is standard for MITOS) and measures of architecture and speed.

Results with precision, recall and F-score are summarized in Table 5. DenseNet is the clear winner in this contest with F-score of 0.927 compared to 0.890 for ESPNet, 0.865 for ResNet and 0.847 for ShuffleNet. Furthermore, results with respect to architecture and speed are summarized in Table 6. Here ResNet is the most efficient with ESPNet a close second.

4. Discussion

While it is the role of the pathologist to make cancer diagnoses and evaluate for important prognostic indicators, such as mitoses, concerning levels of variability have been noted among pathologists (Elmore et al., 2015; Elmore et al., 2017). Variability has been noted both between different pathologists reviewing the same case (inter-observer variability) and within the same pathologist when they are shown the same case on two different occasions, usually with a "wash-out" period between interpretations and they are not told that they are seeing the same cases (intra-observer variability). Clinically, this variability is noted by the submitting clinician if a second opinion is received from another institution. The submitting clinician will not know which opinion is closer to the true biologic nature of the lesion sampled due to the lack of well-established ancillary tests in these circumstances. This places the submitting clinician in the difficult position of discussing variability with the patient, who will likely have associated anxiety of not knowing if their lesion is truly benign or malignant in addition to making the difficult decision of having to decide which treatment option to undergo.

One microscopic parameter that is both helpful to the pathologist in establishing a cancer diagnosis and in assessing prognosis, is the presence or absence of mitotic figures; a microscopically visible nuclear feature closely tied to cellular proliferation. In mitosis a cell divides to form two new cells. Cancer tissue generally has more mitotic activity than normal tissues, and this is assessed by calculation of the mitotic index – the number of cells in mitosis divided by the total number of cells However, measurement of the mitotic index depends on the subjective visual analysis by pathologists who have a hard time both in identifying and also counting mitotic figures and total cell counts (Knezevich et al., 2014). Thus, development of supporting tools that can be more accurate and reproducible would greatly aid clinical care. Machine learning techniques, including CNNs, have shown incredible performance in

Table 5
Performance comparison of ESPNet, DenseNet, ResNet, and ShuffleNet on **MITOS** (Roullier et al., 2010).

Method	ESPNet	DenseNet	ResNet	ShuffleNet
Precision	0.916	**0.939**	0.931	0.968
Recall	0.866	**0.916**	0.807	0.753
F-Score	0.890	**0.927**	0.865	0.847

Table 4
Performance comparison of ESPNet and DenseNet with other approaches on **MITOS** (Roullier et al., 2010) reported in the literature.

Method	ESPNet *(Our trained model)*	DenseNet *(Our trained model)*	Saha et al., 2018	Dodballapur et al., 2019	Li et al., 2018b	López-Tapia et al., 2019	Cireşan et al., 2013 **
Precision	0.916	**0.939***	0.92	0.93	0.854	N/A	0.886
Recall	0.866	**0.916***	0.88	0.80	0.812	N/A	0.70
F-score	0.890	**0.927***	0.90	0.87	0.832	0.826	0.782

* Precision, recall, and F-score of our DenseNet model are higher than other approaches in the literature on the MITOS dataset.
** ICPR12 winner.

Table 6
Architecture, training and inference time comparison of ESPNet, DenseNet, ResNet, and ShuffleNet on **MITOS** (Roullier et al., 2010).

Network	#params (in million)	#blocks (depth)	#channels (width)	Training time*	Inference time*
ESPNet	0.078	16	16 to 64	6 min	8 sec
DenseNet	28.68	161	48 to 2024	19 min	31 sec
ResNet	11.69	12	64 to 512	4 min	6 sec
ShuffleNet	2.28	56	24 to 1024	6 min	11 sec

* Experiments were performed on a 2.10 GHz Intel Xeon Silver 4110 CPU with GeForce GTX 1080 GPU.

visual recognition tasks, and thus have the potential to improve histologic diagnostics, both as aids for pathologists to improve the quality and reproducibility of their diagnoses and in the medical research domain (Mehta et al., 2018a; Ribli et al., 2018; Kermany et al., 2018).

In this work, we trained two CNN methods, ESPNet and DenseNet, as two separate classifiers; both CNNs had high accuracy on our dataset of skin biopsies of invasive melanoma. We further generalized our classifiers to the MITOS breast biopsy dataset and compared our results with the existing state-of-the-art on the MITOS dataset with high accuracy in classifying mitoses (Saha et al., 2018; Cireşan et al., 2013; Chen et al., 2016; Li et al., 2018b; López-Tapia et al., 2019; Dodballapur et al., 2019) and ran experiments with two more state-of-the-art CNNs to make more thorough comparisons. We achieved competitive accuracy on the MITOS dataset compared to the existing state-of-the-art methods.

No study is without limitations, and our research is not an exception. First, both the melanoma dataset and the MITOS dataset (as well as other public digital datasets) make use of less information than a microscopic examination, in which a typical tissue section is 5 μm and on which the pathologist can focus through an infinite number of planes, ensuring all cells of interest are in optimal focus. Secondly, for the public datasets, the use of only two-dimensional images with no recourse to looking at three-dimensional tissue sections makes it difficult to confirm the given diagnoses.

Marking biopsy images is an onerous task and obtaining samples with variation in the dataset is a challenge. To expand our dataset, we generated new samples out of our existing samples with horizontal and vertical mirroring and with rotations of 45, 90, 135 or 225 degrees. However, having samples from more patients would be beneficial for training a precise classifier for mitosis detection.

Given the complex and dense nature of working with biopsy tissue datasets, a significant challenge is posed in developing training sets that reflect the full spectrum of cases seen in clinical practice and also that accurately identify the cellular entity of interest. In our skin cancer work, the cases were carefully selected to represent the full spectrum of skin biopsies obtained in clinical practice and a three-person expert defined consensus diagnosis was used (Elmore et al., 2017). In addition, each case was carefully reviewed by an expert dermatopathologist to identify and mark the individual mitotic figures.

Mitotic activity is an important biomarker that can assist in the diagnosis and may provide prognostic information. However, each biopsy specimen may contain hundreds of thousands of cells, making their identification a significant challenge. We have shown that mitoses can be identified using our machine learning method with high accuracy; thus, this method has the potential of being a powerful diagnostic and prognostic aid to practicing pathologists.

CRediT authorship contribution statement

Shima Nofallah: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft. **Sachin Mehta:** Methodology, Writing - review & editing. **Ezgi Mercan:** Writing - review & editing. **Stevan Knezevich:** Resources. **Caitlin J. May:** Resources. **Donald Weaver:** Resources. **Daniela Witten:** Methodology. **Joann G. Elmore:** Funding acquisition, Supervision, Resources. **Linda Shapiro:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors reported no declarations of interest.

Acknowledgements

This research was supported by NIH grants R01 CA200690, U01CA231782. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

Chen, H., et al., 2016. Mitosis detection in breast cancer histology images via deep cascaded networks. Thirtieth AAAI Conference on Artificial Intelligence.
Cireşan, D.C., et al., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.
Corredor, G., et al., 2018. A watershed and feature-based approach for automated detection of lymphocytes on lung cancer images. In: Medical Imaging 2018: Digital Pathology. International Society for Optics and Photonics.
De Boer, P.-T., et al., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134 (1), 19–67.
Deng, J., et al., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE.
Dodballapur, V., et al., 2019. Mask-Driven Mitosis Detection in Histopathology Images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE.
Elmore, J.G., et al., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama 313 (11), 1122–1132.
Elmore, J.G., et al., 2017. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 357, j2813.
Esteva, A., et al., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.
He, K., et al., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Huang, G., et al., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Irshad, H., Roux, L., Racoceanu, D., 2013. Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE.
Kermany, D.S., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131 e9.
Kingma, D., Adam, B.J., 2014. A Method for Stochastic Optimization arXiv preprint arXiv: 14126980 Cited on: p. 50.
Knezevich, S.R., et al., 2014. Variability in mitotic figures in serial sections of thin melanomas. J. Am. Acad. Dermatol. 71 (6), 1204–1211.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems.
Li, Y., et al., 2018a. Efficient and Accurate Mitosis Detection-A Lightweight RCNN Approach. ICPRAM.
Li, C., et al., 2018b. DeepMitosis: mitosis detection via deep detection, verification and segmentation networks. Med. Image Anal. 45, 121–133.
López-Tapia, S., Aneiros-Fernández, J., de la Blanca, N.P., 2019. A fast pyramidal bayesian model for mitosis detection in whole-slide images. In: European Congress on Digital Pathology. Springer.
Martel, A.L., et al., 2017. An image analysis resource for cancer research: PIIP—pathology image informatics platform for visualization, analysis, and management. Cancer Res. 77 (21), e83–e86.
Mehta, S., et al., 2018a. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.
Mehta, S., et al., 2018b. Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation. Proceedings of the European Conference on Computer Vision (ECCV).
Mercan, E., et al., 2019. Assessment of machine learning of breast pathology structures for automated differentiation of breast Cancer and high-risk proliferative lesions. JAMA Network Open 2 (8) p. e198777-e198777.
Piepkorn, M.W., et al., 2014. The MPATH-Dx reporting schema for melanocytic proliferations and melanoma. J. Am. Acad. Dermatol. 70 (1), 131–141.
Prati, R.C., Batista, G.E., Monard, M.C., 2009. Data Mining with Imbalanced Class Distributions: Concepts and Methods. IICAI.
Ren, S., et al., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems.
Ribli, D., et al., 2018. Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. 8 (1), 4165.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.
Roullier, V., et al., 2010. Mitosis extraction in breast-cancer histopathological whole slide images. In: International Symposium on Visual Computing. Springer.
Roux, L., et al., 2013. Mitosis detection in breast cancer histological images an ICPR 2012 contest. J. Pathol. Inform. 4.
Saha, M., Chakraborty, C., Racoceanu, D., 2018. Efficient deep learning model for mitosis detection using breast histopathology images. Comput. Med. Imaging Graph. 64, 29–40.
Sertel, O., et al., 2009. Computer-aided prognosis of neuroblastoma: detection of mitosis and karyorrhexis cells in digitized histological images. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE.
Simo-Serra, E., et al., 2015. Discriminative learning of deep convolutional feature point descriptors. Proceedings of the IEEE International Conference on Computer Vision.
Society, A.C., 2016. Cancer Facts & Figures. American Cancer Society.
Thompson, J.F., et al., 2011. Prognostic significance of mitotic rate in localized primary cutaneous melanoma: an analysis of patients in the multi-institutional American Joint Committee on Cancer melanoma staging database. J. Clin. Oncol. 29 (16), 2199.
Veta, M., 2016. Tumor Proliferation Assessment Challenge, 2016.
Veta, M., et al., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med. Image Anal. 20 (1), 237–248.
Wang, H., et al., 2014. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. In: Medical Imaging 2014: Digital Pathology. International Society for Optics and Photonics.
Zhang, X., et al., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Zhou, Y., Mao, H., Yi, Z., 2017. Cell mitosis detection using deep neural networks. Knowledge Based Syst. 137, 19–28.