Neural Computing and Applications (2022) 34:8977–8987
https://doi.org/10.1007/s00521-022-06920-3

ORIGINAL ARTICLE

Ovarian cancer detection using optical coherence tomography and convolutional neural networks

David Schwartz[1] · Travis W. Sawyer[1] · Noah Thurston[1] · Jennifer Barton[1] · Gregory Ditzler[1]

Received: 16 June 2021 / Accepted: 4 January 2022 / Published online: 24 January 2022

Abstract

Ovarian cancer has the sixth-largest fatality rate in the United States among all cancers. A non-surgical assay capable of detecting ovarian cancer with acceptable sensitivity and specificity has yet to be developed. However, such a discovery would profoundly impact the pace of the treatment and improvement to patients' quality of life. Achieving such a solution requires high-quality imaging, image processing, and machine learning to support an acceptably robust automated diagnosis. In this work, we propose an automated framework that learns to identify ovarian cancer in transgenic mice from optical coherence tomography (OCT) recordings. Classification is accomplished using a neural network that perceives spatially ordered sequences of tomograms. We present three neural network-based approaches, namely a VGG-supported feed-forward network, a 3D convolutional neural network, and a convolutional LSTM (Long Short-Term Memory) network. Our experimental results show that our models achieve a favorable performance with no manual tuning or feature crafting, despite the challenging noise inherent in OCT images. Specifically, our best performing model, the convolutional LSTM-based neural network, achieves a mean AUC (± standard error) of 0.81 ± 0.037. To the best of the authors' knowledge, no application of machine learning to analyze depth-resolved OCT images of whole ovaries has been documented in the literature. A significant broader impact of this research is the potential transferability of the proposed diagnostic system from transgenic mice to human organs, which would enable medical intervention from early detection of an extremely deadly affliction.

Keywords Deep learning · Optical coherence tomography · Supervised learning

1 Introduction

Cancer is currently the second leading cause of death in the United States, and the number and percentage of people that get cancer in their lifetime have seen an increase in the past 15 years. Ovarian cancer was found to be the 6th most frequent cause of death due to cancer in the U.S [1]. Ovarian cancer is particularly devastating due to its non-specific symptoms, many of which are considered idiopathically harmless when assessed in isolation. The impact of cancer is compounded by the lack of a useful early screening tool, leading to a late-diagnosis rate of 80% [2]. If ovarian cancer is found and can be treated before metastasis, the five-year survival rate is 94% (compared to a baseline of 28% for metastatic cases) [2]. This provides clear evidence for the need for an effective early detection technique.

A non-surgical and high-throughput ovarian cancer screening method would provide a tremendous improvement in quality of life and prognosis. Several imaging techniques have been investigated toward this end. One technique that has shown tremendous promise is optical coherence tomography (OCT). OCT is an interferometric imaging technique that yields depth-resolved, high-resolution images that carry information about the imaged tissue's microstructure. Historically, OCT has been applied with much success to biological imaging in the human eye [3–5], the lung [6, 7], the esophagus [8], the coronary artery [9, 10], and a number of other organs including the ovaries [11–13]. The physical principle of OCT systems is similar to that of ultrasound, except that OCT systems

✉ David Schwartz
dmschwar@email.arizona.edu

[1] University of Arizona, 1230 E Speedway Blvd, Tucson, AZ 85721, USA

Table 1 Neural network optimization parameters

Parameter	VGG	Conv. LSTM	3D CNN
Parameterization of Proposed Models			
Number of learned parameters	23,121,729	36,076,359	1,140,477
Training time per sample (mean ± S.E.)	201 ± 0.96 ms	700 ± 2.62 ms	417 ± 2.55 ms
Mini-batch size	120	120	120
Dropout rate	0.5	N/A	0.5
Training epochs	50	50	50
Batch size	2	2	2
Input normalization	True	False	True
Number of CV replications	10	10	10

measure time-resolved backscattered light instead of sound waves [14]. In particular, OCT images have a wealth of microstructural features in the ovaries, including the stroma, epithelium, and collagen, which show great potential for disease diagnostics and tissue classification [11, 12, 15–18]

One factor stymieing efforts to use OCT for ovarian cancer screening is the optical noise produced by tomographic imaging of the ovaries [19]. Additionally, the data are three-dimensional, subject to scaling challenges, and yield depth-dependent imaging performance. In addition to the characteristic speckle noise, these factors render tomograms extremely challenging for human radiologists and oncologists to diagnose reliably. As a result, advanced computational techniques such as machine learning methods could provide the means to extract quantitative diagnostic information for cancer screening. This manuscript presents our assessment of state-of-the-art neural network-based classification algorithms that solve this task. The results show tremendous promise in machine learning for detecting early tissue changes near the onset of cancer. Our experiments demonstrate that deep VGG-like 3D convolutional neural networks, as well as convolutional LSTMs (Long Short-Term Memory; similar in architecture to those employed previously), achieve high diagnostic accuracy when evaluated on a dataset collected by acquiring optical coherence tomography (OCT) recordings of mouse ovaries in a mouse model of the development of ovarian cancer, introduced in Sect. 4.1 [20, 21].

This manuscript is organized as follows: Section 2 defines our nomenclature. Section 3 summarizes related work. In Sect. 4.1, data acquisition processes are outlined. Section 4.3 exposes our data preprocessing routine. Section 4.4 contains information on the neural network models investigated in this work. In Sect. 5, we present an evaluation of the diagnostic efficacy of these neural networks. Results are interpreted in Sect. 6.

2 Nomenclature

In this section, we introduce variables, parameters and general notation used throughout this work. First, we define a sample, $\mathbf{X}_t$, to be a sequence of OCT images as $\mathbf{X}_t = [\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,j}, \ldots, \mathbf{x}_{t,N}]$, where t indexes the animal sample, N is the number of slices in a sequence, and $\mathbf{x}_{t,j}$ represents the j^{th} image corresponding to the animal indexed by t. This sequence of images is formed by concatenating the subsequences of tomograms of the left and right ovaries. For example, consider the case where $N = 50$. Then each $\mathbf{X}_t$ is represented as 50 sequential images (i.e., 25 images selected from approximately the same depth on each ovary), progressing from the most superficial to the deepest slices. Each animal is assigned a label, $y_t \in \{0, 1\}$, where $y_t = 1$ indicates animal t is predisposed to developing ovarian cancer by 8 weeks of age (i.e., transgenic), while $y_t = 0$ corresponds to wild type (WT) mice. $\hat{y}_t$ denotes predictions of these labels computed by the neural network. In this work, a dataset is a collection of tuples $\mathcal{D} := \{(\mathbf{X}_t, y_t)\}_{t=1}^{T}$.

3 Related work

OCT provides an abundance of information about tissue health. However, quantitatively analyzing three-dimensional OCT data of the ovaries is challenging due to the dimensionality of the data, the depth-dependence, the presence of speckle noise, and the sizeable biological variation inherent to the ovaries. To date, quantitative analyses for OCT images of whole ovaries has been limited to first and second-order statistical techniques such as texture, shape, and frequency analysis [22–24]. These approaches that use OCT imaging technology have shown promise for quantification of tissue changes with the onset of different types of cancer. Despite these quantitative techniques' success, disease detection's sensitivity and specificity could be greatly improved when coupled with more sophisticated machine learning techniques.

Neural networks and related approaches have shown remarkable promise in the context of biological OCT imaging. For example, machine learning has demonstrated utility in assessment for glaucoma [25, 26], retinal diseases [27–30], pulmonary cancer [20], and neurodegenerative and dermatological disease [31, 32]. Other applications of machine learning that have demonstrated success include the modality of intravascular OCT imaging, where neural networks and random forests have been used for atherosclerotic plaque identification [33–35]. Neural networks were also able to detect COVID-19 in pulmonary x-ray imaging [36]. The success of machine learning in cancer diagnosis suggests these techniques in the domain of ovarian tissue imaging could greatly advance the technology toward clinical application. To the best of our knowledge, deep neural networks have not been used to analyze depth-resolved OCT images of whole ovaries.

Machine learning applied to cancer identification has meaningfully benefited from transfer learning [37, 38]. In transfer learning, a model (e.g., neural network) is trained on a task that is not necessarily related to the task on which the model will be evaluated. For example, training a neural network to classify ImageNet data solves a problem quite distinct from that of cancer detection [39]. However, the neural network trained on ImageNet may have learned some useful features from the real scenery that can transfer to cancer classification from OCT imagery. Transfer learning has been successfully used in many tasks and we use transfer learning in this work to boost the performance of our predictive model [38].

4 Methodology

4.1 Data acquisition and imaging

OCT images were collected from a swept source OCT system (OCS1050SS, Thorlabs). The system was set to operate in non-contact mode with a central wavelength of 1040 nm and spectral bandwidth of 80 nm. The axial scan rate was 16 kHz and the power on the sample was measured as 0.36 mW. The system was set to average 4 axial scans, with $11\mu m$ transverse resolution and μm axial resolution in tissue. The total imaging volume was 4 mm × 4 mm lateral, and 2 mm deep. The digital images are 750 × 752 × 512 pixels (pixel size of approximately 5 μm × 5 μm). The image volume was exported as a series of 2D images (or slices).

The OCT data in this work were initially curated for automated segmentation algorithms [19], and 3D texture analysis [15]. Ideally, the sequence of OCT images can be concatenated in a third dimension to visualize the 3D structure of an organ. Unfortunately, a major challenge with OCT data is that the noise statistics associated with optical backscattering vary with organ depth and presumably tissue health. For example, common irregularities attributed to variations in tissue density, optical absorption characteristics, and concentration of scatterers impeded early attempts at quantitative analysis of optical coherence tomograms [19]. To ameliorate the impacts of these inconsistencies, we propose a Gaussian blur during preprocessing to smooth the images. All remaining computation to counteract any deleterious effects of optical noise resides in the neural network classifiers described in Sect. 4.4.

4.2 Mouse model

The image data were collected from a transgenic mouse model (TgMISIIR-TAg) in which females spontaneously develop bilateral epithelial ovarian cancer [40, 41]. All TAg positive (TAg+) TgMISIIR-TAg female mice develop bilateral epithelial ovarian cancer, with invasive tumors in the ovaries evident in nearly all mice by eight weeks of age. Sixteen mice were sacrificed at eight weeks for imaging (eight TAg+, eight wild type) and explanted organs were imaged using the OCT system. Details on mouse breeding protocol, and surgical explantation can be found in previous publications [15, 19, 42]. All imaged tissue was analyzed via immunohistochemistry and evaluated by a pathologist for the presence and extent of tumors, which is determined via cell morphology and presence of the TAg protein. This process provides a thorough validation that the TAg+ mice exhibit ovarian cancer by eight weeks of age. Further details on the histological analysis can be found in a previous work by Sawyer et al. [43].

4.3 Data preprocessing

The tomography imagery consists of 750 × 752 pixel images (see Fig. 1a), with pixel intensities in [0, 255]. We

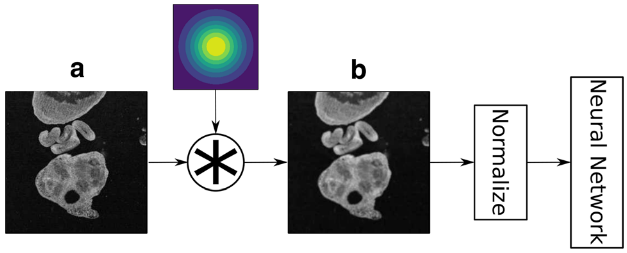


Fig. 1 An illustration of the sequence of preprocessing steps under consideration: **a** depicts the raw image of slice 100 of the left ovary of animal 3767, **b** shows the result of convolving this with a 2-dimensional Gaussian kernel

perform a sequence of preprocessing transformations to render these images useful to neural networks. Figure 1 highlights the pipeline of preprocessing operations. First, we rescale pixel intensities linearly to the interval, $[-1, 1]$. A prior study of this OCT dataset revealed that speckle noise inherent in the medium significantly confounds automatic segmentation systems' efforts to isolate ovarian tissue [19]. In order to reduce this noise and improve perceptibility of the images we pass each image through a Gaussian filter (with a standard deviation of 1) to produce Fig. 1b. The Gaussian filter was empirically shown to mitigate the effects of the noise (e.g., compared to median, low-pass, or anisotropic filters) for the segmentation task studied in [19]. At the final stage of preprocessing, we standardize each image (i.e., normalize by calculating the pixel-wise mean and standard deviation of intensity from selected training data, then subtracting this mean from each pixel value and dividing by the empirical standard deviation). After preprocessing each image, the next phase in our cancer detection framework is to train a deep neural network to perform the task of classification of sequences of OCT images.

4.4 Classification model

Convolutional neural networks (CNNs) are a class of artificial neural network consisting of banks of neurons whose output states (which can be thought of as pixels in a visual analogy) are computed as the convolution of an input signal with the filter learned by the given bank of neurons. CNNs were introduced as a solution to the problem of handwritten digit classification and have since been applied nearly ubiquitously in computer vision tasks [20, 44, 45]. VGG is a remarkably deep CNN that achieves near state-of-the-art performance on challenging image classification tasks, including medical image analysis [45, 46]. As shown in Fig. 2, the VGG-based model consists of sequential convolutional layers, represented as yellow volumes, and pooling operations (i.e., down-sampling via pixel aggregation), represented as orange volumes. The primary convolutional block is composed of two two-dimensional serially connected spatial convolution layers outputting 64 channels into a max-pooling layer. This pattern is repeated in subsequent blocks as illustrated, successively doubling the number of channels in a convolutional layer's output until the final two convolutional blocks (each of which outputs 512 channels). Blocks 3-5 all have three serially connected convolutional layers. Block 5 feeds into a fully connected neural network (i.e., the re-encoding layer), which feeds into a second fully connected layer (i.e., the decoding layer) with a single sigmoidally activated neuron (as opposed to the rectified exponential nonlinearity defined in Eq. 2), which ensures that the final layer solves the classification problem by effectively performing logistic regression on the penultimate layer's encoding of the OCT imagery. The output of the decoding layer indicates an estimated likelihood of each class (WT vs transgenic) for the given sample. Unlike the original implementation of the VGG network, which uses ReLU nonlinearities, all convolutional layers in our model signal with rectified exponential activation functions.

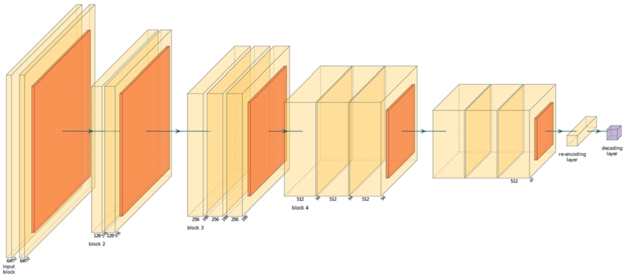


Fig. 2 A graphical depiction of the VGG architecture investigated in this work. The primary convolutional block is composed of two 2D serially connected spatial convolution layers (represented in yellow) outputting 64 channels. These feed into max-pooling layers represented in orange. This pattern is repeated in subsequent blocks as illustrated, successively doubling the number of channels in a convolutional layer's output until the final two convolutional blocks (each of which outputs 512 channels). Blocks 3-5 all have three serially connected convolutional layers. Block 5 feeds into a fully connected neural network (the re-encoding layer), which feeds into a second fully connected layer (the decoding layer) with a single sigmoid neuron (as opposed to the rectified exponential nonlinearity). The output of the decoding layer indicates an estimated likelihood of each class (WT vs transgenic) for the given sample (Color figure online)

We initialize the VGG sub-network with weights learned on Imagenet, which contains photographic images, to leverage transfer knowledge [47]. Transfer learning is the approach taken in many computer vision applications where a pre-trained deep neural network is first optimized on an unrelated dataset, in which there is an abundance of labeled data [38, 48, 49], before being fine-tuned on the task-relevant dataset. The pre-trained network provides feature maps (i.e., nonlinear feature extractors) that are learned from Imagenet. Once the network is pre-trained, we fine-tune the network on the OCT data described in the previous section. Despite any suspected disparity between the generation of imagery of natural scenes compared with that of biological tissues (e.g., melanoma dermoscopy compared with Imagenet), transfer learning has shown to be beneficial in other neural network-based medical image tasks [45] and applications [50–52]. VGG minimizes the cross-entropy between the probability distribution underlying ground truth ($\{y_t\}_t : \sim p_y$) and the distribution of decisions decoded from the output of the model, denoted as $p_{\hat{y}}$ (each implicitly conditioned on the data, $\{\mathbf{X}_t\}$):

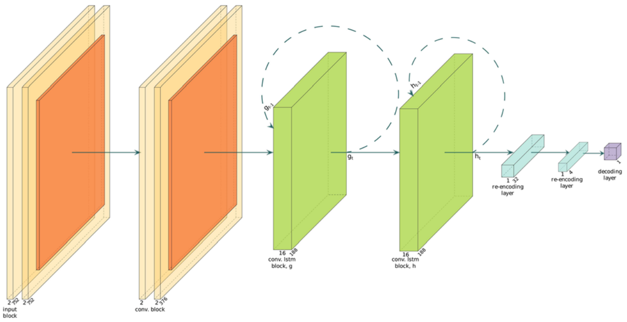


Fig. 3 A graphical depiction of the convolutional LSTM architecture investigated in this work. The input and convolutional blocks are feed-forward layers consisting of a pair of 2-channel, 2-dimensional spatial convolutions (yellow) which feed into max-pooling aggregation layers (orange). 16-channel convolutional LSTMs comprise the central layers (green), g and h, with feedback connections represented by dashed arrows connecting each iteration's output (e.g., g) to the subsequent iterations input (g_{t-1}). For each slice, t, h_t, is relayed to a sequence of fully connected feed-forward layers (turquoise) that re-encode it for classification in the decoding layer (purple) (Color figure online)

$$\mathcal{L} = -\frac{1}{M}\sum_{t=1}^{M}[y_t \log \hat{y}_t + (1 - y_t) \log (1 - \hat{y}_t)], \tag{1}$$

where M is the number of images in the training dataset. $\mathcal{L}$ is the cross-entropy loss function and can be thought of as an empirically estimated KL divergence between the distributions of ground truth and predictions. Minimizing this loss function tends to drive a model toward maximizing the information it encodes about its training dataset [53, 54].

Long short-term memories (LSTMs) are a class of recurrent neural networks that learn temporal dependencies in data in recurrent connections gated by their constituent LSTM cells [55]. Recently a new class of convolutional neural network equipped with the feedback connections and gates that distinguish LSTMs from earlier recurrent architectures has demonstrated favorable performance in precipitation forecasting [21] and anomaly detection in video [56]. Inspired by these results, we also use a convolutional LSTM that learns spatial correlations inherent in 3D tomography data as temporal relationships in its training data [57]. A convolutional LSTM is depicted graphically in Fig. 3. The input and convolutional blocks consist of 2-channel, 2-dimensional spatial convolutions that feed into max-pooling layers. As with conventional CNNs, the max-pooling layers downsample each channel in their input to half resolution [45]. There are 16-channel convolutional LSTMs that comprise the next layers (shown in green), g_t and h_t, with feedback connections represented by dashed arrows. The convolutional LSTM layers instantiate architectures described by Xingjian et al. and initialize intermediate states, g and h, as zeros [21]. The second convolutional LSTM layer's output, h, is relayed to a sequence of fully connected feed-forward layers that re-encode h for classification by the decoding layer. Other than the model's output, which uses sigmoid activation functions, every other neuron in this model uses the rectified version of the exponential linear activation [58]. For completeness, the rectified exponential linear activation function is explicitly defined as

$$\Psi(x) = \begin{cases} \exp(x) - 1, & x < 0, \\ x, & x \in [0, 1], \\ 1, & x > 1 \end{cases} \tag{2}$$

We also experiment with another neural network model, namely 3D CNNs. The 3D CNNs are an extension of convolutional layers and model a spatial dimension along which imaging data are arranged. 3D CNNs have found success in applications ranging from human pose estimation [59] to medical image analysis [20, 60]. A 3D CNN is implemented nearly identically to 2D CNNs, differing only in the number of dimensions over which convolutions are evaluated. A 2D CNN filter evaluates a single 2D convolution of an image with a 2D kernel (i.e., an image). In contrast, a 3D CNN filter convolves sequences of images with 3D kernels (i.e., volumes). These 3D-CNN architectures consist of three feed-forward subnetworks: (a) a sequentially distributed (i.e., "TimeDistributed" in the nomenclature of TensorFlow) 2D convolutional neural network, (b) a 3D-CNN, in which the third dimension is formed by ordering elements of each sequence, $\mathbf{x} \in \mathbf{X}_t$, and (c) a multilayer perceptron responsible for estimating the likelihood that each $\mathbf{x}_{t,k} \in \mathbf{X}_t$ belongs to a transgenic animal. As shown in Fig. 4, the primary convolutional block contains feed-forward layers consisting of 2D spatial

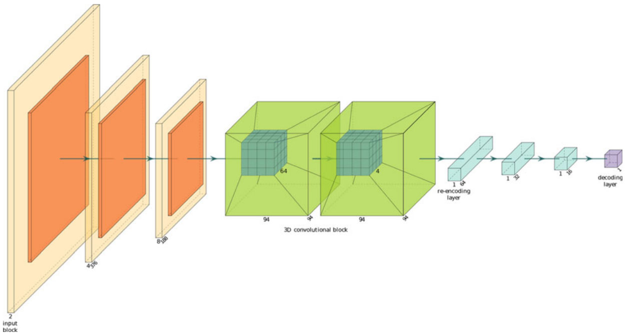


Fig. 4 A graphical depiction of the 3D CNNs investigated in this work. The primary convolutional block contains feed-forward layers consisting of 2D spatial convolutions (yellow) which feed into max-pooling aggregation layers (orange). These are organized with 2, 4, and 8 channels (i.e., 2D filters) in the first, second, and third convolutional layers, respectively. A 64 channel 3D CNN connected to a 4 channel 3D CNN make up the central layers (green). For each slice, t, h_t, is relayed to a sequence of fully connected feed-forward layers (turquoise) that re-encode it for classification in the decoding layer (purple) (Color figure online)

convolutions (yellow) which feed into max-pooling aggregation layers (orange). These are organized with 2, 4, and 8 channels (i.e., 2D filters) in the first, second, and third convolutional layers, respectively. A 64 channel 3D CNN connected to a four channel 3D CNN makes up the central layers (green). For each slice, *t*, the outputs of the 3D convolutional block are connected to a sequence of fully connected feed-forward layers (turquoise) that re-encode them for classification in the decoding layer (purple).

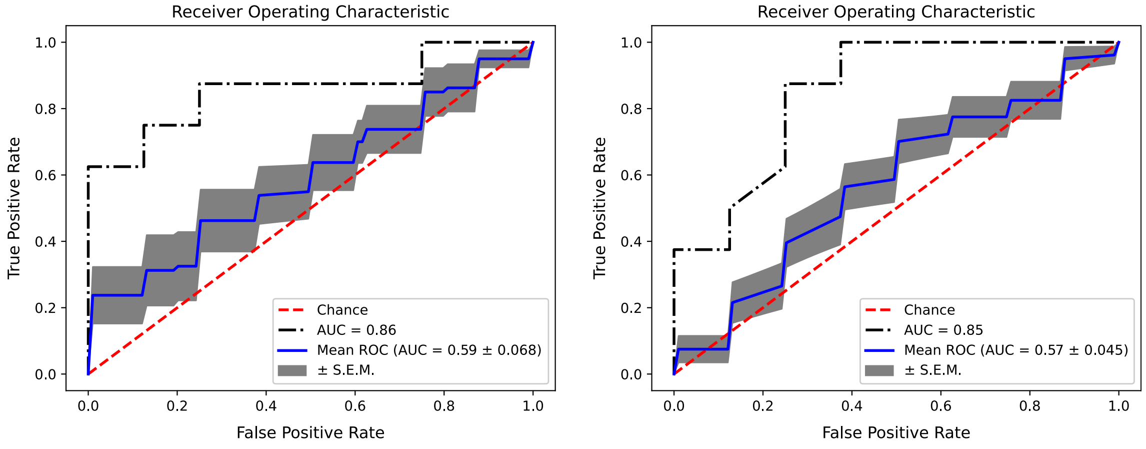
5 Results

In this section, we present an empirical analysis of the VGG, Convolutional LSTM and 3D-CNN on the OCT dataset described in Sect. 4.1. These comparisons also allow us to determine each algorithm's strengths and weaknesses for the task of cancer detection. We begin our discussion of the results and findings by describing the experimental paradigm and model parameterizations studied.

5.1 Model parameterizations

This subsection offers an in-depth description of model parameterizations and configurations. We use dropout at the connections from VGG to the re-encoding layer [61], through which the outputs of VGG's penultimate layer are randomly and dynamically zeroed out during training. Stochastically setting neurons' outputs to zero during training typically reduces training time while guiding optimization away from deep local minima in the loss function. Weights and biases in the first two layers (i.e., those belonging to the input block shown in Fig. 2) are held constant throughout the learning routine. Fixing these parameters to the values optimized on Imagenet reduces training time (by decreasing the number of variable parameters) and has no significant effect on average and peak AUC. A marginal (but insignificant) enhancement in mean (and peak) AUC can be seen by comparing the ROCs summarized in Fig. 6a with those in Fig. 5, for which parameters of all layers are variables learned in optimization. Optimization is regularized by augmenting $\mathcal{L}$ (in Eq. 1) with a penalization (weighted by a factor of 0.0005) of the L_2 norm of the weights learned in the re-encoding layer. The weights and biases of this model are optimized by the "Nadam" routine [62], an extension of the popular Adam optimization algorithm that incorporates Nesterov momentum to increase the rate of convergence of the optimization process. The learning rate is initially set to 0.001, and is adapted as a function of gradients of $\mathcal{L}$. In contrast to the 3D CNNs and convolutional LSTMs, for which Batch Normalization (BN) [63] (the process of normalizing the outputs of intermediate layers of a neural network) was necessary in order to stabilize learning, incorporating BN between the intermediate layers of our implementation of VGG did not seem to affect performance metrics assessed here significantly.

Distinct from 3D CNN- and VGG-based models, the convolutional LSTM model proposed perceives OCT imagery that has not been normalized as described in Sect. 4.3. Also, unlike the VGG- and 3D CNN-based models, dropout is not used while training our LSTM-based models. Cross-entropy loss is optimized and



(a) Primary parameters fixed

(b) All parameters variable

Fig. 5 A comparison of summaries of ROCs achieved by training **a** an instance of VGG in which the weights and biases of the two primary layers, which were optimized on Imagenet, are fixed during learning with **b** an instance of VGG in which weights and biases are initialized randomly and remain variable throughout optimization. A marginal but likely insignificant improvement due to transfer learning is evident in the differences in geometries of the peak ROCs. However, the improvement in area under ROC is a small fraction of the standard error of the mean (i.e., the shaded region)

regularized by the L_2 norm of each layers' weights. In contrast to the VGG-based and 3D CNN models considered in this work, convolutional LSTMs are optimized using the Adadelta algorithm [64], which is empirically a more stable (and computationally parsimonious) choice for this architecture and dataset. Batch normalization was applied to the inputs of the intermediate layers of the convolutional LSTM block during training to stabilize the estimation of gradients of $\mathcal{L}$ with respect to the parameters belonging to these layers [63]. Batch normalization was found to accelerate learning and improve generalization performance. The learning rate is initially set to 0.001, and is adapted as a function of gradients of $\mathcal{L}$ as proposed by Zeiler [64].

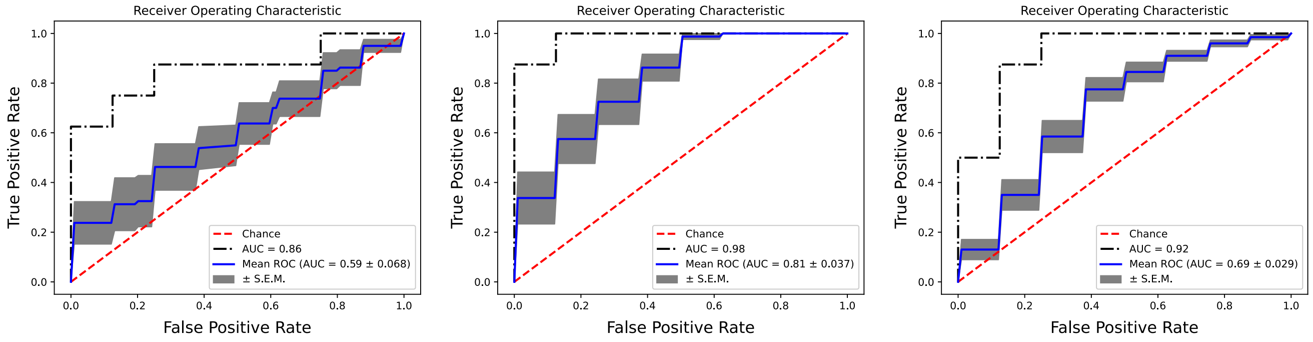
The 3D CNN model proposed is the only model whose performance was experimentally shown to benefit from the normalization of OCT imagery described in Sect. 4.3. As in our implementations of the VGG- and LSTM-based models, we train 3D CNNs subject to dropout rate of 50% while minimizing cross-entropy loss regularized with the L_2 norm of the weights. The weights of this architecture are optimized using the Nadam algorithm [65]. Without BN applied to the inputs of the intermediate layers of the 3D convolutional block, the performance of the 3D CNN models assessed here is critically impaired. Exactly as with the VGG-based models, the learning rate is initially set to 0.001 and is adapted as a function of gradients of $\mathcal{L}$.

5.2 Cross-validation experiment

Generalization performance is the most important set of statistics that we are interested in understanding. The experiments seek to measure the performance on data never observed in the past during training time. We devised a leave-one-out cross-validation (CV) experiment to test our models' ability to generalize to unseen data. Newly initialized models are trained and validated on a subset of the complete set of tomography sequences before evaluating the hold-out animal's sequence. Specifically, for each animal in our dataset, we perform one fold of CV. Within each fold, the remaining 15 animals are divided among a singleton containing an animal whose label equals the test animal and seven disjoint sets containing two animals (one transgenic and one WT). Our models are trained on sequential mini-batches corresponding to these seven stratified subsets. Validation subsets are selected as the next mini-batch that the model will train on to cope with physical memory constraints and maximize the number of mini-batches whose training is validated by an as-yet-unseen subset of samples. In the final mini-batch, which must be validated on already-seen data, the validation set is chosen to be the training data exposed in the first mini-batch. Validating models on unseen data during training serves an additional role in mitigating catastrophic forgetting. Catastrophic forgetting is a phenomenon observed in learning, where a neural network forgets previously learned knowledge as it is exposed to new information [66]. We also use early stopping during training to reduce the risk of over-training. Early stopping is implemented by halting training on batches for which further training does not improve the loss on the validation subset. Our training routine ensures that each model learns from at most a single positive and negative sample in each mini-batch. After training is complete, the model in question is evaluated on the held-out test sample.

5.3 Performance results

Figure 6 compares our models' diagnostic efficacies (i.e., their ability to predict the occurrence of ovarian cancer from OCT images in the transgenic mouse model described in 4.1). Efficacy is assessed by the Receiver Operating Characteristic (ROC) curve, which plots true positive rate (i.e., $\Pr(\hat{y} = 1|y = 1)$ = sensitivity) against false positive rate (i.e., $\Pr(\hat{y} = 1|y = 0) = 1 -$ specificity) [67]. The red dashed line shows the result of random prediction (i.e., uniformly random guessing), which is the worst performance that a classifier can achieve. We also report the area under the ROC, which approximates the probability that a given model will rank the likelihood of a positive sample higher than that of a randomly chosen negative sample. These statistics are summarized in Table 2. The mean ROC curves shown are interpolated from the CV experiment described in Sect. 5.2. The convolutional LSTM achieved maximum AUC, showing a marginal improvement of only 0.06 over the 3D CNN (see Table 2 for the peak AUC and average AUCs with the standard error). In contrast, the VGG-based model is significantly underperformed, only achieving a maximum AUC of 0.86. Interestingly, the VGG model achieved the worst AUC despite requiring the greatest amount of time to train. The 3D CNN and convolutional LSTM incur similar time costs, but complete a single training epoch in less than half the time required for the VGG-based model to do the same. The empirically most powerful classifiers evaluated in this work, the convolutional LSTMs that achieved a peak AUC of 0.98, committed only a single false positive (and no other errors). Based on these results, the convolutional LSTM shows a tremendous amount of promise for ovarian cancer detection from OCT imagery.



(a) VGG **(b)** Convolutional LSTM **(c)** 3D-CNN

Fig. 6 Receiver operating characteristic (ROC) curves computed by interpolating the functional mean ROC from recordings of replications of the aforementioned CV experiment for **a** VGG, **b** a convolutional LSTM, and **c** a 3D-CNN corresponding to the parameterizations outlined Sect. 5.1. The shaded error region shown is within one standard error of the mean ROC curve. The dashed red curve (for which true positive rate is equal to false positive rate) is an idealized ROC corresponding to classifying by random chance (i.e., uniformly random guessing). The dashed black curve is the ROC that achieved the maximum area enclosed below among all replications of the CV experiment

Table 2 Peak and average AUCs achieved over ten replications of the leave-one-out cross-validation experiment described in Sect. 5.2, summarized from the results shown in Fig. 6

Model	Peak AUC	Mean AUC ± SE
Areas under ROC		
VGG	0.86	0.59 ± 0.068
Conv. LSTM	0.98	0.81 ± 0.037
3D-CNN	0.92	0.69 ± 0.029

The standard error is measured on a 90% confidence interval

6 Discussion

6.1 Conclusions

This work's contributions form a critical first step toward an automatic OCT-based human ovarian cancer diagnostic system. The proposed classifiers learn and adapt abstract representations of tomograms conducive to detecting radiographic signatures of ovarian cancer in OCT imagery without manual feature selection. Results presented here show that (to the extent of the limits imposed by the dataset), highly discriminatory classifiers that can be expected to generalize to unseen data can be evolved. Moreover, their incurrence of very few misclassifications is replicable across multiple runs of the leave-one-out cross-validation program.

To the best of the authors' knowledge, this is the first demonstration of a proof-of-concept model for cancer detection using depth-resolved OCT recordings of ovaries, which has been shown to be a challenging medium on which to base inferences of genotype in both OCT and widefield fluorescence [19, 68–70]. A recent approach to OCT-based ovarian cancer detection using a generalized linear model classifier showed promising results for detecting malignant (vs. normal) ovarian tissue [71]. However, that effort is distinct from ours in that they imaged biopsies of ovarian tissue using full field OCT and performed classification on hand-crafted features developed by human analysis of ovarian OCT data. In contrast, our methodology learns features maps from the training data, and our proposed neural networks are benchmarked on depth-resolved OCT recordings of intact ovaries.

6.2 Future work

With an admittedly small dataset, consisting of only 16 total animals, future experimentation with the proposed classifiers must involve validation on a larger dataset, which would enable larger cross-validation experiments where many animals are held out for testing on each fold. We emphasize that to the extent of the limits imposed by the dataset analyzed in this work, the cross-validation results presented are exclusively the results of generalization performance (i.e., all testing is performed on samples that do not appear in the training subset). However, this procedure suffers from the limitation of only assessing a single test animal in the test phase. A larger dataset enabling a larger cross-validation experiment would allow us to draw stronger conclusions on diagnostic efficacy with reduced uncertainty. Additionally, a larger collection of mouse OCT imagery may provide valuable information to be leveraged in a transfer learning experiment when eventually adapting the models for human subjects. An incredibly useful extension of the models presented here is a quantitative method to identify features and regions in the OCT imagery that leads to a neural network's decision (i.e., the specific region in the OCT image where the tumor is present). These regions could provide medical practitioners insight into the uncertainty of a neural network's prediction. For example, consider the case in which an artificial occlusion (e.g., an implanted medical device) or an interferometric artifact partially obscures (or mimics) a radiographic signature of cancer. Unless such occlusions are sufficiently common throughout the classifier's training dataset, it is unlikely that the neural network has learned to accurately identify the signatures of ovarian cancer in the presence of the occlusion. Therefore, health care providers may decide that the result warrants further consideration, perhaps in conjunction with other assays (e.g., collecting serum to identify or exclude the possible presence of biomarkers that indicate the progression of ovarian cancer [72]).

6.3 Broader impacts

Perhaps this work's most profound broader impact lies in potentially dramatically improving the likelihood of detecting ovarian cancer in patients before metastasis throughout the peritoneal cavity, which would radically improve treatment outcomes. That our models were trained and evaluated on a transgenic mouse model of ovarian cancer development begs a central question: *to what extent does a neural network from our work transfer to OCT data collected from humans* Given the difficulty of collecting such data, developing an even larger dataset of mouse ovary tomograms may prove advantageous if the knowledge learned from the mouse model is relevant for analyzing human ovarian OCT data.

Funding This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant No. DGE-1143953; NSF CAREER #1943552; Department of Energy #DE-NA0003946; National Institutes of Health under National Cancer Institute #1R01CA195723; and the shared resources of the University of Arizona Cancer Center #3P30CA023074. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

Data availability Data will be madeavailable upon request.

Code availability Source code can be found at https://github.com/dmschwar/OCT-based-OCD.

Declarations

Conflicts of interest The authors have no conflicts of interest to report.

Ethical approval All experiments were performed per NIH guidelines, and protocols were approved by the University of Arizona Institutional Animal Care and Use Committee under protocol 06-183.

References

1. US Cancer Statistics Working Group, "US cancer statistics data visualizations tool, based on november 2018 submission data (1999-2016): US Department of Health and Human services, Centers for Disease Control and Prevention and National Cancer Institute," *Centers for Disease Control and Prevention and National Cancer Institute*, 2019
2. Buys SS, Partridge E, Black A, Johnson CC, Lamerato L, Isaacs C, Reding DJ, Greenlee RT, Yokochi LA, Kessel B et al (2011) Effect of screening on ovarian cancer mortality: the prostate, lung, colorectal and ovarian (plco) cancer screening randomized controlled trial. Jama 305(22):2295–2303
3. Swanson Ea, Izatt Ja, Hee MR, Huang D, Lin CP, Schuman JS, Puliafito Ca, Fujimoto JG (1993) In vivo retinal imaging by optical coherence tomography. Opt Lett 18(21):1864–6
4. Hee MR, Izatt JA, Swanson EA, Huang D, Schuman JS, Lin CP, Puliafito CA, Fujimoto JG (1995) Optical coherence tomography of the human retina. Arch Ophthalmol 113(3):325
5. Abràmoff M, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. IEEE Rev Biomed Eng 1(3):169–208
6. Tsuboi M, Hayashi A, Ikeda N, Honda H, Kato Y, Ichinose S, Kato H (2005) Optical coherence tomography in the diagnosis of bronchial lesions. Lung Cancer 49(3):387–394
7. Otte S, Otte C, Schlaefer A, Wittig L, Hüttmann G, Dromann D, Zeli A (2013) "OCT A-Scan based lung tumor tissue classification with Bidirectional Long Short Term Memory networks," In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6
8. Lightdale CJ (2013) Optical coherence tomography in Barrett's esophagus. Gastrointest Endosc Clin N Am 23(3):549–563
9. Ferrante G, Presbitero P, Whitbourn R, Barlis P (2013) "Current applications of optical coherence tomography for coronary intervention"
10. Abdolmanafi A, Duong L, Dahdah N, Cheriet F (2017) Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography. Biomed Opt Express 8(2):1203
11. Hariri LP, Liebmann ER, Marion SL, Hoyer PB, Davis JR, Brewer MA, Barton JK (2010) Simultaneous optical coherence tomography and laser induced fluorescence imaging in rat model of ovarian carcinogenesis. Cancer Biol Ther 10(5):438–447
12. Wang T (2015) An overview of optical coherence tomography for ovarian tissue imaging and characterization. Wiley Interdiscip Rev Nanomed Nanobiotechnol 7(1):1–16
13. Drexler W, Liu M, Kumar A, Kamali T, Unterhuber A, Leitgeb RA (2014) Optical coherence tomography today: speed, contrast, and multimodality. J Biomed Opt 19(7):071412
14. Schmitt J (1999) Optical Coherence Tomography (OCT): a review. IEEE J Sel Top Quantum Electron 5(4):1205–1215
15. Sawyer T, Chandra S, Rice P, Koevary J, Barton J (2018) Three-dimensional texture analysis of optical coherence tomography images of ovarian tissue. Phys Med Biol 63:23
16. Welge WA, DeMarco AT, Watson JM, Rice PS, Barton JK, Kupinski MA (2014) Diagnostic potential of multimodal imaging of ovarian tissue using optical coherence tomography and second-harmonic generation microscopy. J Med Imag 1(2):025501
17. Brewer Ma, Utzinger U, Barton JK, Hoying JB, Kirkpatrick ND, Brands WR, Davis JR, Hunt K, Stevens SJ, Gmitro AF (2004) Imaging of the ovary. Technol Cancer Res Treat 3(6):617–627
18. Watanabe Y, Takakura K, Kurotani R, Abe H, Atanabe YUW, Akakura KEIT, Urotani REK (2015) Optical coherence tomography imaging for analysis of follicular development in ovarian tissue. App Opt 54(19):6111

19. Sawyer TW, Rice PF, Sawyer DM, Koevary JW, Barton JK (2018) Evaluation of segmentation algorithms for optical coherence tomography images of ovarian tissue. Diagn Treat Dis Breast Reprod Syst IV 10472:1047204
20. Alakwaa W, Nassef M, Badr A (2017) Lung cancer detection and classification with 3d convolutional neural network (3d-cnn). Lung Cancer 8(8):409
21. S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo (2015) "Convolutional lstm network: A machine learning approach for precipitation nowcasting," In *Advances in neural information processing systems*, 802–810
22. Gossage KW, Tkaczyk TS, Rodriguez JJ, Barton JK (2003) Texture analysis of optical coherence tomography images: feasibility for tissue classification. J Biomed Opt 8(3):570–575
23. Miller P, Astley S (1992) Classification of breast tissue by texture analysis. Image Vis Comput 10(5):277–282
24. Mostaço-Guidolin LB, Ko AC-T, Wang F, Xiang B, Hewko M, Tian G, Major A, Shiomi M, Sowa MG (2013) Collagen morphology and texture analysis: from statistics to classification. Sci Rep 3(1):2190
25. Ran AR, Tham CC, Chan PP, Cheng C-Y, Tham Y-C, Rim TH, Cheung CY (2020) "Deep learning in glaucoma with optical coherence tomography: a review," *Eye*
26. Burgansky-Eliash Z, Wollstein G, Chu T, Ramsey JD, Glymour C, Noecker RJ, Ishikawa H, Schuman JS (2005) Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. Invest Ophthalmol Vis Sci 46(11):4147–52
27. Yanagihara RT, Lee CS, Ting DSW, Lee AY (2020) Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review. Trans Vision Sci Technol 9:11–2
28. Rahimy E (2018) Deep learning applications in ophthalmology. Current Opin Ophthalmol 29(3):254–260
29. Ditzler G, Bouaynaya N, Fathallah Shaykh HM (2019) Sparse kalman filtering for time-varying networks. BMC BioData Min 12:1–14
30. Ditzler G, Bouaynaya N, Shterenberg R (2018) AKRON: an algorithm for approximating sparse kernel reconstruction. Signal Process 144:265–270
31. Johri A, Tripathi A (2019) *et al.*, "Parkinson disease detection using deep neural networks," In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–4, IEEE
32. Yasir R, Rahman MA, Ahmed N (2014) "Dermatological disease detection using image processing and artificial neural network," In *8th International Conference on Electrical and Computer Engineering*, pp. 687–690, IEEE
33. Lee J, Prabhu D, Kolluru C, Gharaibeh Y, Zimin VN, Bezerra HG, Wilson DL (2019) Automated plaque characterization using deep learning on coronary intravascular optical coherence tomographic images. Biomed Opt Express 10:6497–6515, 11
34. Lee J, Prabhu D, Kolluru C, Gharaibeh Y, Zimin VN, Dallan LAP, Bezerra HG, Wilson DL (2020) Fully automated plaque characterization in intravascular OCT images using hybrid convolutional and lumen morphology features. Sci Rep 10:2596
35. He C, Li Z, Wang J, Huang Y, Yin Y, Li Z (2020) "Atherosclerotic Plaque Tissue Characterization: An OCT-Based Machine Learning Algorithm With ex vivo Validation "
36. Nour M, Cömert Z, Polat K (2020) A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization. Appl Soft Comput 97:106580
37. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3(1):1–40
38. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
39. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252
40. Connolly DC, Bao R, Nikitin AY, Stephens KC, Poole TW, Hua X, Harris SS, Vanderhyden BC, Hamilton TC (2003) Female mice chimeric for expression of the simian virus 40 TAg under control of the MISIIR promoter develop epithelial ovarian cancer. Cancer Res. 63(6):1389–1397
41. Quinn BA, Xiao F, Bickel L, Martin L, Hua X, Klein-Szanto A, Connolly DC (2010) Development of a syngeneic mouse model of epithelial ovarian cancer. J Ovarian Res 3(1):24
42. Watson JM, Rice PF, Marion SL, Bentley DL, Brewer MA, Utzinger U, Hoyer PB, Barton JK (2011) Multi-modality optical imaging of ovarian cancer in a post-menopausal mouse model. In: Advanced biomedical and clinical diagnostic systems IX, vol 7890. International Society for Optics and Photonics, p 78900W
43. Sawyer T, Koevary J, Rice F, Howard C, Austin O, Connolly D, Cai K, Barton J (2019) Quantification of multiphoton and fluorescence images of reproductive tissues from a mouse ovarian cancer model shows promise for early disease detection. J Biomed Opt 24(9):096010
44. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
45. Jaworek-Korjakowska J, Kleczek P, Gorgon M (2019) "Melanoma thickness prediction based on convolutional neural network with vgg-19 model transfer learning," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0
46. Simonyan K, Zisserman A (2014) "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556
47. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) "Imagenet: A large-scale hierarchical image database," In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee
48. Xiang EW, Cao B, Hu DH, Yang Q (2010) Bridging domains using world wide knowledge for transfer learning. IEEE Trans Knowl Data Eng 22(6):770–783
49. Pan S, Tsang I, Kwok J, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210
50. Schweikert G, Widmer C, Schölkopf B, Rätsch G (2008) An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In: NIPS, vol 8. Citeseer, pp 1433–1440
51. Ahmed A, Yu K, Xu W, Gong Y, Xing E (2008) "Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks," In *European Conference on Computer Vision*, pp. 69–82
52. Guo J, Liang Z, Scribner E, Ditzler G, Bouaynaya N, Fathallah-Shaykh H (2018) "Nonlinear brain tumor model estimation with long short-term memory neural networks," In *IEEE/INNS International Joint Conference on Neural Networks*
53. Zhang Z, Sabuncu M (2018) "Generalized cross entropy loss for training deep neural networks with noisy labels," In *Advances in neural information processing systems*, pp. 8778–8788
54. Wang Y, Ma X, Chen Z, Luo Y, Yi J, Bailey J (2019) "Symmetric cross entropy for robust learning with noisy labels," In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 322–330
55. Gers FA, Schmidhuber J, Cummins F (1999) "Learning to forget: Continual prediction with lstm," *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*

56. Luo W, Liu W, Gao S (2017) "Remembering history with convolutional lstm for anomaly detection," In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 439–444, IEEE
57. Graves A, Fernández S, Schmidhuber J (2005) "Bidirectional lstm networks for improved phoneme classification and recognition," In *International Conference on Artificial Neural Networks*, pp. 799–804, Springer
58. Clevert D-A, Unterthiner T, Hochreiter S (2015) "Fast and accurate deep network learning by exponential linear units (elus)," arXiv preprint arXiv:1511.07289
59. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C (2017) "Monocular 3d human pose estimation in the wild using improved cnn supervision," In *2017 international conference on 3D vision (3DV)*, pp. 506–516, IEEE
60. Ren X, Xiang L, Nie D, Shao Y, Zhang H, Shen D, Wang Q (2018) Interleaved 3d-cnn s for joint segmentation of small-volume structures in head and neck ct images. Med Phys 45(5):2063–2075
61. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
62. Dozat T (2016) "Incorporating nesterov momentum into adam," *International Conference on Learning Representations (ICLR)*
63. Santurkar S, Tsipras D, Ilyas A, Madry A (2018) "How does batch normalization help optimization," In *Advances in Neural Information Processing Systems*, pp. 2483–2493
64. Zeile MD (2012) "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701
65. Ruder S (2016) "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747
66. Grossberg S (1988) Nonlinear neural networks: principles, mechanisms, and architectures. Neural Netw 1(1):17–61
67. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874
68. Sawyer TW, Koevary JW, Howard CC, Austin OJ, Rice PF, Hutchens GV, Chambers SK, Connolly DC, Barton JK (2020) Fluorescence and multiphoton imaging for tissue characterization of a model of postmenopausal ovarian cancer. Lasers Surg Med 52(10):993–1009
69. Sawyer TW, Rice FF, Koevary JW, Connolly DC, Cai KQ, Barton JK (2019) In vivo multiphoton imaging of an ovarian cancer mouse model. Dis Breast Reprod Syst V 10856:1085605
70. Sawyer TW, Chandra S, Rice PF, Koevary JW, Barton JK (2018) Three-dimensional texture analysis of optical coherence tomography images of ovarian tissue. Phys Med Biol 63(23):235020
71. Nandy S, Sanders M, Zhu Q (2016) Classification and analysis of human ovarian tissue using full field optical coherence tomography. Biomed Opt Express 7(12):5182–5187
72. Zhang Z, Bast RC, Yu Y, Li J, Sokoll LJ, Rai AJ, Rosenzweig JM, Cameron B, Wang YY, Meng X-Y et al (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res 64(16):5882–5890

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.