# Automated Pneumothorax Diagnosis using Deep Neural Networks*

Tony Lindsey[1,2](✉), Rebecca Lee[2], Ronald Grisell[3], Saul Vega[3], and Sena Veazey[3]

[1] Department of Biomedical Informatics, Stanford University, Palo Alto, CA
[2] NASA Ames Research Center, Mountain View, CA
{antonia.e.lindsey, rebecca.l.lee-1}@nasa.gov
[3] US Army Institute of Surgical Research, San Antonio, TX
{ronald.d.grisell2.ctr, saul.j.vega.ctr,
sena.r.veazey.ctr}@mail.mil

**Abstract.** Thoracic ultrasound can provide information leading to rapid diagnosis of pneumothorax with improved accuracy over the standard physical examination and with higher sensitivity than anteroposterior chest radiography. However, the clinical interpretation of a patient medical image is highly operator dependent. Furthermore, remote environments, such as the battlefield or deep-space exploration, may lack expertise for diagnosing certain pathologies. We have developed an automated image interpretation pipeline for the analysis of thoracic ultrasound data and the classification of pneumothorax events to provide decision support in such situations. Our pipeline consists of image preprocessing, data augmentation, and deep learning architectures for medical diagnosis. In this work, we demonstrate that robust, accurate interpretation of chest images and video can be achieved using deep neural networks. A number of novel image processing techniques were employed to achieve this result. Affine transformations were applied for data augmentation. Hyperparameters were optimized for learning rate, dropout regularization, batch size, and epoch iteration by a sequential model-based Bayesian approach. In addition, we utilized pretrained architectures, applying transfer learning and fine-tuning techniques to fully connected layers. Our pipeline yielded binary classification validation accuracies of $98.3\%$ for M-mode images and $99.8\%$ with B-mode video frames.

**Keywords:** Deep learning · Pneumothorax classification · Ultrasound · Transfer learning · Bayesian optimization

## 1 Motivation

Thoracic ultrasound is a noninvasive, readily-available imaging modality that supplements clinical examination in the evaluation of chest pathologies involving the pleural cavity [6]. In particular, radiologists have identified multiple sonographic artifacts indicative of pneumothorax (PTX), such as sliding lung absence, reverberation, bar code pattern and transition point presence [1, 7]. The expertise available for diagnosis and treatment of PTX may be curtailed in austere or remote locations, such as the

* Supported by the Space Technology Mission Directorate.

battlefield or aboard a deep space exploration vehicle. Due to the limited resources in these environments, artificial intelligence can play a significant role in augmenting clinically-relevant and interpretable medical patient diagnosis. A common occurrence on the battlefield is penetrating or blunt force thoracic injury that impairs the airways and may induce collapsed lung. Although PTX pathology is rare among astronauts, acute hypobaric decompression exposure is a plausible risk factor in microgravity and would present a significant challenge in both diagnosis and immediate treatment. Therefore, reliable interpretation capability for this life threatening condition is imperative for celeritous intervention. We hypothesized that machine learning can be used for accurate, early diagnosis of PTX in traumatic injuries, and accordingly developed an ultrasound medical imaging platform for thoracic pulmonary injury diagnosis. The objective of this study was to build and assess an automated computer model that provides near real-time binary PTX diagnosis of porcine pulmonary ultrasound images. The model's effectiveness was quantified by analyzing performance metrics on the train, validation and test sets.

An intelligent clinical decision support system is a powerful tool that aids clinical management of patient care and treatment for potentially life-threatening injuries as well as evaluating affected areas following medical procedures. Machine learning algorithms trained to distinguish pulmonary feature signs indicative of PTX were developed and examined. These algorithms, when applied to previously unseen test images, exhibited relatively high statistical performance metrics of sensitivity, specificity and positive predictive value. Our foremost algorithm is equipped to work with sonographic M-mode images and B-mode video frames of normal and pathological pulmonary function.

## 2 Dataset

Porcine clinical ultrasound data sources of pulmonary health were supplied by the US Army Institute of Surgical Research. Ground truth categorical binary labels for PTX pathology associated with $404$ M-mode ($209$ bmp, $195$ jpg) images and $420$ B-mode mp4 video clips were provided to build automated medical diagnosis models. Baseline classification accuracy statistics as computed by the iFAST computerized assistant were $97.5\%$ for M-mode images and $84.7\%$ for B-mode video [8]. A Sonosite M-Turbo ultrasound machine captured all images and video loops. Cine-videos were $5$ seconds in duration with a frequency of $40$ frames per second. The associated linear transducer monitored intercostal space $2$ for each subject and acquired images utilizing settings: mechanical index $0.7$, probe depth $4$cm, and soft tissue thermal index $0.1$.

## 3 Methods

In recent years, deep learning has become a topic of much discussion and research due to its impressive pattern recognition discernment on large multi-class data sets, such as ImageNet. Transfer learning has mitigated computation time, improved accuracy and enhanced development of robust deep learning models. Consequently, computer-aided diagnosis via deep learning is now feasible, despite a lack of large medical database prevalence. We have developed a complete software pipeline for the medical diagnosis

of PTX cases from ultrasound image products using convolutional neural networks. In addition to transfer learning, we have utilized various image preprocessing techniques, data augmentation, fine-tuning, and Bayesian optimization.

### 3.1 Data Retrieval

A total of $420$ B-mode videos and $404$ M-mode still images from eight female Yorkshire porcine models (*Sus scrofa*) with and without PTX, used in a previous study [8], were acquired for our experiments. The images were then split into $80\%$ train and $20\%$ test sets.

### 3.2 Preprocessing

Ultrasound images and videos contain artifacts, such as text, lines, tick marks and granular speckle noise. Such manifestations are detrimental for accurate image classification; thus, we developed a digital image processing module to filter such uninformative structures from the ultrasound data prior to developing a learning model. Fig. 1 illustrates steps taken to properly clean the medical images.

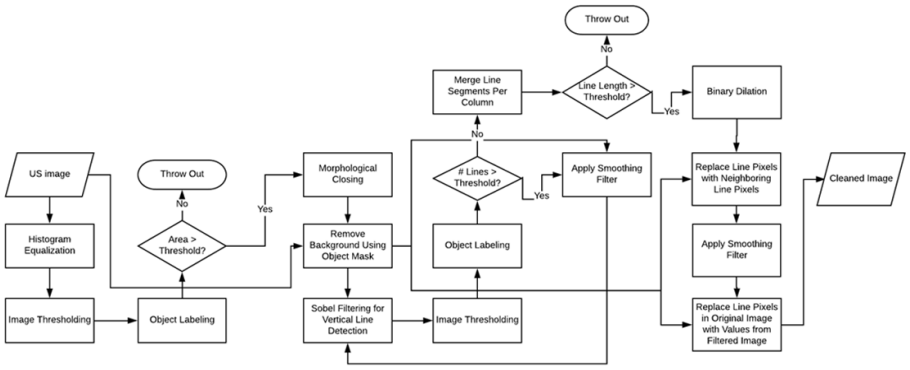


**Fig. 1:** Image preprocessing pipeline module component.

The initial step involved removing image frame background structures. This was accomplished by thresholding the image after histogram equalization was performed. The threshold was determined based on the histogram of intensity values taken from the histogram equalized image. Subsequently, morphological opening, closing, and hole-filling were employed in order to separate neighboring structures and to ensure complete capture of relevant structures. After this was accomplished, we sought to blend vertical line artifacts into the data. Vertical lines were localized using the x-direction Sobel filter. The line pixels were then replaced with neighboring image pixels and finally smoothed with a Gaussian filter.

### 3.3 Data Augmentation

The data augmentation module increased the number of images using affine transformations to improve network localization capability and generalized modeling. Augmentation of images that preserve collinearity and distance ratios was performed prior to model training. Each transform resulted in one additional output generated per image with arguments determining type of augmentation. Flips (horizontal, vertical), angled rotations, translational pixel shifts, regional zoom, random Gaussian noise and blurring by various amounts were implemented. Categorical parity was achieved by supplementing with a 3:1 ratio of negative to positive M-mode generated images. Our module extracted a 1.7:1 ratio of negative to positive frames from B-mode video to reach label parity. Finally, contrast-limited adaptive histogram equalization was applied to all images and frames for enhanced detection of subtle features.

### 3.4 CNN Architecture

Pretrained convolutional neural networks (CNN) are models that have been trained using a large dataset, e.g. ImageNet contains one million images with over a thousand categories. The resulting weighted connections from such a pretrained CNN were utilized to accelerate and transfer learned features with activations available in the penultimate fully connected layer. This particular layer was trained with our porcine dataset for the canonical case. Several deep neural network models were examined to determine optimal architecture for our application. A 16-layers deep model developed by Oxford University's Visual Geometry Group (VGG16) consistently recorded higher diagnostic accuracy than alternative architectures examined. The network consists of $3 \times 3$ convolutional layers stacked in increasing depth while reducing volume size by max pooling. Then two fully connected layers, each with $4,096$ nodes, are followed by a softmax classifier. Increased convolution layers and improved utilization of internal network computing resources allow the network to learn deeper features. For example, the first layer might learn only edges while the deepest layer learns to interpret transition patterns differentiating movement at the pleural lines, such as seashore sign, a normal lung feature. The network contains convolution blocks with activation on the top layer that defines complex functional mappings between inputs and response variables, followed by batch normalization after each convolutional layer.

The max pooling sample-based discretization process was performed with kernel size $3 \times 3$ and stride 2. The network was then flattened to one dimension after the final convolutional block. Dropout of network layers was performed until reaching the dense five node output layer, which uses a softmax activation function to compute the probability of classification labels. Exponential and leaky rectified linear unit activation was applied with gradient value $0.01$ to mitigate dead neuron bottlenecks during backpropagation. The network also used convolutional layer $L_2$ regularization to reduce model overfitting, binary cross-entropy computed error loss, and the Xavier method for initializing weights so that neuron activation functions begin in unsaturated regions. The inclusion of batch normalization improved validation set PTX classification accuracy on average $1.6\%$ across both model modes.

### 3.5 Transfer Learning

Transfer learning based approaches were executed using VGG19, ResNet50 and VGG16 architectures pretrained with weights updated based on ImageNet visual database training. In order to achieve the transfer learning scenario, the last fully connected layer was removed followed by treating the remaining network components as a fixed feature extractor for the new train dataset [5]. The technique retains initial pretrained model weights and extracts image features via a final network layer. Additionally, further "fine-tuning" experiments were performed by extending backpropagation to the last four layers. Due to overfitting concerns, only four higher-level layer dimensions of the network were fine-tuned. Our experiments revealed that fine-tuning yielded improved performance over transfer learning alone.

### 3.6 Bayesian Optimization

A deep neural network's effectiveness is influenced by "higher-level" prior distribution properties of the model, such as complexity and learning rate. The optimal selection of these hyperparameters can be framed as a model validation loss minimization problem. Bayesian optimization is a probabilistic model-based approach for finding the minimum of any objective function that returns a real-value metric, such as CNN validation error with respect to hundreds of model architectures and hyperparameter choices. The approach has been applied to feed-forward computer vision models with greater efficiency than manual, random, or grid search in terms of better overall test set performance and decreased optimization time [3]. In our experiments, we optimized dropout rate, learn-

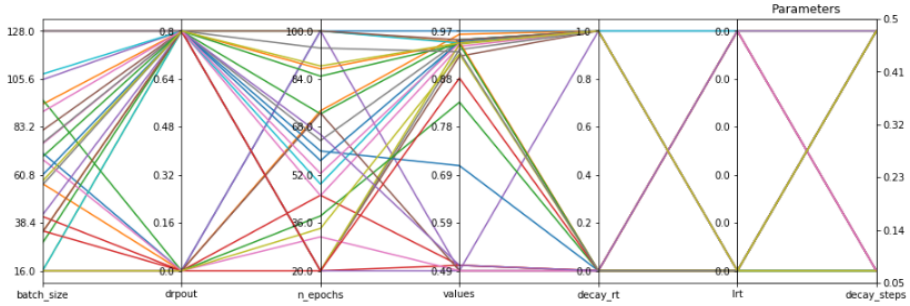


**Fig. 2:** Parallel coordinate multivariate visualization example across 7 dimensions for VGG16 model. Optimized configuration occurs at data cluster points: batch_size=71, drpout=0.0, n_epochs=100, (accuracy) values=0.97, decay_rt=1.0, lrt=1e-05, decay_steps=0.05.

ing rate, decay rate, batch size, training epochs, and decay step size hyperparameters using the GPyOpt Python open-source library package [2].

After defining a search space for the optimization process, a posterior distribution function that best describes the objective function to optimize was constructed. As the number of observations grows, the posterior distribution improves and the algorithm

becomes more certain of which regions in parameter space are worth exploring (see Fig. 2). A Gaussian process model with integrated expected improvement acquisition was used and initially explored 10 random points for the first model fit. This number was double the default value, but necessary since we are explored a complex and noisy 7 dimensional hyperparameter space. Consequently, the Gaussian process model fit required more time using CPU resources to complete than building and optimizing the neural network with a single GPU. Moreover, sampling many random points initially ameliorates the risk of becoming trapped in a local minima. The model explored a maximum of 100 points following the initial parameter sampling results (see Table 1).

| M-mode Gaussian Process Hyperparameter Optimization | | | | | | | | | B-mode Gaussian Process Hyperparameter Optimization | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pretrained Model | Dropout | Learning Rate | Decay Rate | Batch Size | Epochs | Loss | Accuracy | Solver | Pretrained Model | Dropout | Learning Rate | Decay Rate | Batch Size | Epochs | Loss | Accuracy | Solver |
| VGG16 | 0.2825 | 1.00e-03 | 0.9500 | 16 | 100 | 0.0886 | 0.9700 | SGD | VGG16 | 0.3335 | 5.56e-04 | 0.4474 | 16 | 30 | 0.4450 | 0.8057 | SGD |
| VGG16 | 0.2455 | 9.13e-04 | 0.8507 | 32 | 50 | 0.0534 | 0.9743 | Adam | VGG16 | 0.2708 | 2.14e-04 | 0.9431 | 64 | 20 | 0.0313 | 0.9903 | Adam |
| VGG16 | 0.3819 | 3.27e-04 | 0.6191 | 16 | 30 | 0.0654 | 0.9764 | RMSProp | VGG16 | 0.5313 | 1.73e-04 | 0.9474 | 32 | 20 | 0.0437 | 0.9844 | RMSProp |
| VGG19 | 0.1370 | 1.00e-03 | 0.9250 | 16 | 20 | 0.1486 | 0.9422 | SGD | VGG19 | 0.0000 | 9.72e-04 | 1.0000 | 32 | 50 | 0.1861 | 0.9296 | SGD |
| VGG19 | 0.0933 | 2.04e-04 | 0.5969 | 16 | 20 | 0.1045 | 0.9672 | Adam | VGG19 | 0.0492 | 1.60e-04 | 0.7151 | 16 | 20 | 0.0717 | 0.9778 | Adam |
| VGG19 | 0.5960 | 7.50e-04 | 0.8100 | 16 | 100 | 0.1106 | 0.9759 | RMSProp | VGG19 | 0.4488 | 2.47e-04 | 0.9895 | 64 | 20 | 0.0588 | 0.9749 | RMSProp |
| ResNet50 | 0.2790 | 9.89e-04 | 0.9782 | 32 | 100 | 0.4837 | 0.7195 | SGD | ResNet50 | 0.5427 | 1.00e-03 | 0.7377 | 64 | 20 | 0.6706 | 0.6311 | SGD |
| ResNet50 | 0.0640 | 1.00e-03 | 0.8750 | 128 | 50 | 0.3077 | 0.8608 | Adam | ResNet50 | 0.7102 | 2.70e-04 | 0.9366 | 64 | 100 | 0.4778 | 0.7464 | Adam |
| ResNet50 | 0.5789 | 6.59e-04 | 0.8640 | 16 | 50 | 0.3790 | 0.8051 | RMSProp | ResNet50 | 0.6979 | 2.22e-04 | 0.9591 | 32 | 30 | 0.4743 | 0.7561 | RMSProp |

**Table 1:** Gaussian process Bayesian optimization results for 3 evaluated models. Dataset labels were negative (normal lung) and positive (PTX). VGG16 models with stochastic gradient descent solvers (yellow highlight) produced the highest optimized accuracy.

# 4 Experiments and Results

## 4.1 Model Generation

The acquired medical imaging data products were partitioned into train, validation and test sets based using manifest supplied ground truth information. Train and validation sets were augmented using affine transformations that created synthetic images. The M-mode data partitioning resulted in $1,868$ train images, $467$ validation images, and $81$ images for test. B-mode video was partitioned as $16,212$ train frames, $4,053$ validation frames, and $1,013$ images for test. The held-out test subsets were disjoint and analyzed only once by the trained models. The images were cropped to area size $224 \times 224$ and used as input data by a VGG16 architecture previously trained for generic classification tasks on Imagenet visual database. The model was then implemented with a high-level neural network API called Keras, running on the TensorFlow library backend for numerical computation using data flow graphs. An NVIDIA Tesla K80 accelerator hardware device with 12 GiB of GPU memory powered the training and a form of early stopping influenced estimation of optimal test set model epoch. Successive iterations guided construction of more complex model architectures fine-tuned for improved diagnostic interpretation of our pulmonary sonographic datasets.

### 4.2 Binary model classification

Three candidate model architectures were evaluated as binary diagnostic classifiers using porcine PTX medical images. The model validation performance was compared with baseline iFAST logistic regression classifier statistics. The VGG16 model achieved the best overall prediction accuracy. iFAST baseline statistical parameter results were outperformed for both M-mode images and B-mode video frames. Previously discussed image preprocessing, data augmentation, hyperparameter optimization and transfer learning techniques were used as a pipeline process for model generation. The results successfully achieved published state-of-the-art accuracy levels (see Table 2).

| Validation Data Set Statistics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Modality | Model | Accuracy | 95% CI | Sensitivity | Specificity | PPV | NPV | Kappa |
| B-mode | VGG16 | 0.9978 | (0.9958, 0.9999) | 0.9990 | 0.9965 | 0.9966 | 0.9990 | 0.9956 |
| B-mode | iFAST | 0.8465 | (0.8076, 0.8803) | 0.8566 | 0.8258 | 0.9102 | 0.7365 | 0.6617 |
| M-mode | VGG16 | 0.9829 | (0.9665, 0.9926) | 0.9957 | 0.9701 | 0.9707 | 0.9956 | 0.9657 |
| M-mode | iFAST | 0.9753 | (0.9551, 0.9881) | 0.9818 | 0.9618 | 0.9818 | 0.9618 | 0.9436 |

**Table 2:** B-mode: 4053 frames, 2028 true positives, 2 false negatives, 7 false positives, 2016 true negatives. M-mode: 467 images, 232 true positives, 1 false negative, 7 false positives, 227 true negatives.

## 5 Discussion and Future Directions

The VGG16 CNN models recorded a higher ratio of false positives to false negatives for both ultrasound modalities. Lichtenstein found that absence of lung sliding alone is very sensitive for PTX, but not specific in ICU patients due to large numbers of false positives [4]. Porcine subjects that are critically ill or exhibiting pulmonary contusions may cause similar interference and plausibly explain the observation.

There are multiple directions that can be pursued to improve upon our results. Currently, several image processing techniques for contrast enhancement are being evaluated with the intent of improving precision and medical imaging feature discrimination. Our dataset was class imbalanced, a characteristic that negatively affects CNN classifier accuracy. Consequently, we plan to assess generative adversarial networks as an augmentation tool for restoring categorical parity and supplying images of superior quality. Our classifiers achieved $100\%$ accuracy on test set images for both ultrasound modalities. However, more rigorous analysis of the models with larger datasets including human pulmonary structures is necessary and scheduled as a future activity.

## 6 Conclusion

In this paper we presented a fully automatic processing pipeline of thoracic ultrasound for PTX pathology classification. Data retrieval and preprocessing modules acquired ultrasound image products, removed artifacts and employed adaptive histogram equalization for improved image contrast. An initial sparse dataset was synthetically enhanced

with pathology-preserving affine image transformations. Pretrained model weight utilization together with retraining selected fully connected layers improved generalizability and accelerated training time for PTX feature learning. Error analysis revealed that learning rate, optimizer type and image preprocessing were the greatest contributors to overall improved pipeline processing element performance. Bayesian optimization determined an optimal hyperparameter model configuration which outperformed random search according to our experiments.

## 7 Acknowledgements

We acknowledge Jose Salinas, PhD, chief of the Clinical Decision Support and Automation research at the US Army Institute of Surgical Research, which provided laboratory and biomedical device equipment support. We also acknowledge Maria Serio-Melvin, MSN, and Army Col. Shawn Nessen, DO, FACS, for their invaluable trauma patient clinical insights.

## References

1. Alrajhi, K., Woo, M., Vaillancourt, C.: Test characteristics of ultrasonography for the detection of pneumothorax: a systematic review and meta-analysis. CHEST Journal **141**, 703–708 (2012)
2. Authors, T.G.: GPyOpt: A bayesian optimization framework in python. `http://github.com/SheffieldML/GPyOpt` (2016)
3. Bergstra, J., Yamins, D., Cox, D.D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on Machine Learning **28**(1), 115–123 (Jun 2013)
4. Lichtenstein, D., Menu, Y.: A bedside ultrasound sign ruling out pneumothorax in the critically ill. Lung sliding. Chest **108**(5), 1345–8 (Nov 1995)
5. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1717–1724 (2014)
6. Qureshi, N., Rahman, N., Gleeson, F.: Thoracic ultrasound in the diagnosis of malignant pleural effusion. Thorax Journal **64**, 139–143 (2009)
7. Stone, M.B.: Ultrasound diagnosis of traumatic pneumothorax. J Emerg Trauma Shock **1**(1), 19–20 (Jan 2008)
8. Summers, S.M., Chin, E.J., April, M.D., Grisell, R.D., Lospinoso, J.A., Kheirabadi, B.S., Salinas, J., Blackbourne, L.H.: Diagnostic accuracy of a novel software technology for detecting pneumothorax in a porcine model. Am J Emerg Med **35**(9), 1285–1290 (Sep 2017)