EBioMedicine 62 (2020) 103094

Contents lists available at ScienceDirect

EBioMedicine

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

Research paper

Automatic discovery of clinically interpretable imaging biomarkers for *Mycobacterium tuberculosis* supersusceptibility using deep learning

Thomas E. Tavolara[a], M. Khalid Khan Niazi[a,*], Melanie Ginese[b], Cesar Piedra-Mora[c], Daniel M. Gatti[d], Gillian Beamer[b], Metin N. Gurcan[a]

[a] Center for Biomedical Informatics, Wake Forest School of Medicine, 486 Patterson Avenue, Winston-Salem, NC 27101, United States
[b] Department of Infectious Disease and Global Health, Tufts University Cummings School of Veterinary Medicine, 200 Westboro Rd., North Grafton, MA 01536, United States
[c] Department of Biomedical Sciences, Tufts University Cummings School of Veterinary Medicine, 200 Westboro Rd., North Grafton, MA 01536, United States
[d] The College of the Atlantic, 105 Eden Street, Bar Harbor, ME 04609, United States

ARTICLE INFO

Article History:
Received 20 May 2020
Revised 9 October 2020
Accepted 12 October 2020
Available online xxx

Keywords:
Granuloma
Lung
Diversity outbred mice
Multiple instance learning
Biomarkers
Tuberculosis
Machine learning
Mice

ABSTRACT

Background: Identifying which individuals will develop tuberculosis (TB) remains an unresolved problem due to few animal models and computational approaches that effectively address its heterogeneity. To meet these shortcomings, we show that Diversity Outbred (DO) mice reflect human-like genetic diversity and develop human-like lung granulomas when infected with *Mycobacterium tuberculosis* (*M.tb*) .
Methods: Following *M.tb* infection, a "supersusceptible" phenotype develops in approximately one-third of DO mice characterized by rapid morbidity and mortality within 8 weeks. These supersusceptible DO mice develop lung granulomas patterns akin to humans. This led us to utilize deep learning to identify supersusceptibility from hematoxylin & eosin (H&E) lung tissue sections utilizing only clinical outcomes (supersusceptible or not-supersusceptible) as labels.
Findings: The proposed machine learning model diagnosed supersusceptibility with high accuracy (91.50 ± 4.68%) compared to two expert pathologists using H&E stained lung sections (94.95% and 94.58%). Two non-experts used the imaging biomarker to diagnose supersusceptibility with high accuracy (88.25% and 87.95%) and agreement (96.00%). A board-certified veterinary pathologist (GB) examined the imaging biomarker and determined the model was making diagnostic decisions using a form of granuloma necrosis (karyorrhectic and pyknotic nuclear debris). This was corroborated by one other board-certified veterinary pathologist. Finally, the imaging biomarker was quantified, providing a novel means to convert visual patterns within granulomas to data suitable for statistical analyses.
Implications: Overall, our results have translatable implication to improve our understanding of TB and also to the broader field of computational pathology in which clinical outcomes alone can drive automatic identification of interpretable imaging biomarkers, knowledge discovery, and validation of existing clinical biomarkers.
Funding: National Institutes of Health and American Lung Association.

1. Introduction

Tuberculosis (TB) is an important global disease. Over 2 billion people are currently infected with *Mycobacterium tuberculosis* (*M.tb*), the bacterium that causes TB. In 2018, 10 million people were diagnosed with TB and 1.5 million people died, surpassing mortality due to HIV/AIDS [1]. More than 4000 deaths occur per day due to TB. When infected with *M.tb,* those susceptible develop lung disease, called active pulmonary TB. Clinical symptoms are fever, cough, progressive weight loss and emaciation, due to lung inflammation, necrosis, and cavitation. Death occurs in 40-70% of untreated cases [2,3]. Fortunately, most (90%) humans are resistant to *M.tb* and, if infected, survive with latent *M.tb* infection (LTBI) for decades. Although a small fraction of LTBI cases transition to active TB due to acquired immunodeficiency, diabetes, and old age [4], most patients who progress to active pulmonary TB have no known risk factors. Additional forms of TB are recognized clinically (e.g. fulminant, miliary, subclinical, and incipient) and are part of the spectrum of human responses to *M.tb*. There are no consensus lung or blood biomarkers readily available to diagnose active pulmonary TB, and examination

* Corresponding author.
E-mail address: mniazi@wakehealth.edu (M.K.K. Niazi).

https://doi.org/10.1016/j.ebiom.2020.103094

Research in context

Evidence before this study

We performed literature searches for publications in English without date restrictions. We searched PubMed for "tuberculosis AND (mouse model) AND (granuloma necrosis)" on April 27th 2020. This retrieved 116 publications: 15 reviews and 101 primary research articles. Twenty-six primary articles reported granuloma necrosis: 24 with inbred or gene-deleted mice, and 2 with humanized mice or inbred mice with human transgenes. No primary publications except our own used Diversity Outbred mice. No primary research articles applied artificial intelligence to H&E lung granulomas. We searched PubMed for (tuberculosis AND ("imaging biomarker") on April 27th 2020. This retrieved 3 publications, all on PET or CT scans of human TB patients and none on histology images. We searched PubMed and arXiv for the term "(multiple instance learning) AND (histology OR pathology OR histopathology)" on September 18, 2019 without date restriction and limited to English articles. This resulted in a total of 35 articles from PubMed and 24 articles from arXiv. Six of these results applied deep multiple instance learning to predict some diagnostic label using histopathology images. All studies evaluated their methodology on cancer histology slides, including colon, prostate, basal cell, and breast. Five studies used weak labels for training of their models, and one required regions of interest. Five studies applied their method to a two-class problem. Four used an attention-based mechanism for pooling multiple instances into a slide-level decision. All studies mentioned interpretability, but only one study emphasized this and verified it with a pathologist. Further, no study compared model performance to pathologist performance or non-expert performance, verified model-identified key instances with current clinical practices, quantified model-identified key instances, or recognized the potential for their model to discover novel imaging biomarkers (in which tissue-level annotations are impossible).

Added value of this study

Attention-based deep learning can automatically discover clinically relevant histopathology-based biomarkers and can quantify these features. The latter (quantification) is impossible for human pathologists to perform and is a major strength of this study as visual information can be extracted and quantified for statistical analyses. Further, in supersusceptibility to *M.tb* infection, both our model and non-experts exceeded pathologist performance using model-identified imaging biomarkers.

Implications of all the available evidence

Deep learning in pathology is stifled by its requirement for meticulously curated ground truth, inefficiency in processing large images (like in histopathology), and its "black-box" nature. Our experiments with *M.tb*-infected mice implicate a model that automatically identifies interpretable imaging biomarkers for disease using only a diagnostic label. Not only does our proposed machine learning model meet these shortcomings of deep learning, it also more broadly paves a path for the discovery of new interpretable imaging biomarkers and validation of current practices in clinical pathology.

of lung tissue in sick patients is performed by in-life scans, after surgical removal or at autopsy. As few animal models develop human-like lung granulomas, there is a need for additional animal models of TB to improve translational relevance of experimental findings.

To address needs for a better mouse model of TB, we use Diversity Outbred (DO) mice. Each DO mouse's genome is a heterozygous mosaic of DNA inherited from the 8 founder strains (Supplemental Fig. 1). The population was created by breeding together 5 *Mus musculus* spp *domesticus* strains (A/J; C57BL/6J; 129S1/SvlmJ; NOD/ShiLtJ; NZO/HILtJ) and 3 wild-derived *Mus musculus* strains (CAST/EiJ; PWK/PhJ; WSB/EiJ) [5,6]. The genetic diversity of the DO population rivals the genetic diversity of the human population and this has been exploited to understand the genetic basis of disease [6–14]. When infected virulent *M.tb,* DO responses better emulate human forms of TB than C57BL/6, BALB/c, CBA/J, or C3HFeB/HeJ inbred strains [2,15–19]. Analogous forms of TB in humans and DO mice are shown below (Table 1).

Following infection with a low dose of aerosolized *M.tb* bacilli, DO mice develop wide phenotype ranges in survival, weight change, lung granuloma morphotypes, acquired immunity, and innate inflammatory responses [20–22] that are not observed in inbred strains including the related Collaborative Cross recombinant inbred lines [5,6,19,22–30]. Like humans, weight loss in *M.tb*-infected DO mice reflects metabolic signatures of poorly regulated inflammation (unpublished). Also like humans, biomarker signatures from lungs and/or serum better discriminate TB disease forms in DO mice than single biomarkers [23]. Since those initial studies, we are now generating larger data sets to find accurate protein biomarkers for diagnosis and to generate testable mechanistic hypothesis. Finally, we have applied image analysis and deep learning models to automatically detect histologic features of *M.tb*-infected lungs, including granulomas, cell-poor caseous necrosis, lymphocytic cuffs, macrophage-rich regions, neutrophil-rich regions, normal lung tissue, and acid-fast stained *M.tb* [21,23,31–33].

Our extensive deep learning methods to automatically extract information from lungs of *M.tb*-infected DO mice are limited by three key factors, reflected in digital pathology more generally. First, the vast majority of deep learning models require *strong* labels [34]. Strong labels can be thought of as any label that can be delineated in an image (such as nuclei or tissue layers), whereas *weak* labels can describe an image more generally (e.g. a morphological diagnosis applied to a slide). Acquisition of strong labels or manual annotations for digital histology images is a barrier for deep learning in computational pathology [35]. In our problem – diagnosis of supersusceptibility to *M.tb* – manual annotation for strong labels is not an option, as there are no consensus histopathological features or specific cell types indicative of supersusceptibility. Second, deep learning models do not efficiently process the amount of data in digital histology images, on the order trillions of pixels. In other domains, large images can be resized [36–41]. However, in computational pathology, resizing results in information loss including individual cells, locations, and tissue-level microanatomy (like looking at low-magnification). As an option to resizing, tiles (small images cropped from a large digital images) are often substituted for the whole image, but this approach still requires manual annotations (i.e. strong labels) and is computationally expensive to process hundreds of thousands of tiles per tissue section. Third, how deep learning models make decisions for a histology side is not interpretable by humans and therefore difficult to trust. This 'black-box' nature of computational pathology limits its acceptance in biomedical research and medicine, as both scientists and clinicians wish to know how decisions are made before the information is used to inform a biological mechanism or to make an informed clinical decision. Tools such as class activation mapping (CAM) [42], Grad-CAM [43], and Grad-CAM++ [44] can highlight which parts of an image contribute to what the deep learning model "sees" but cannot be applied in computational pathology for diagnosis, as digitized tissue sections are too large.

To overcome these limitations, we implement an attention-based multiple instance learning (MIL) [45] model to identify supersusceptible DO mice using hematoxylin and eosin (H&E) stained lung tissue

Table 1
Analogous TB form in humans and DO mice.

Humans (survival)	Fulminant TB (weeks)	Pulmonary TB (months/years)	Incipient TB (years)	Latent TB infection (years/decades)	Early clearance (normal lifespan)
DO mice (survival)	Supersusceptible (<8 weeks)	Susceptible (12–20 weeks)	Resistant (>20 weeks)	Superresistant (unknown)	Not yet observed (unknown)

sections. Unlike conventional deep learning in computational pathology, the proposed machine learning model requires only weak labels, can efficiently process the large digital histopathology images without resizing, is interpretable, and automatically identifies regions in H&E slides that contribute to its clinical label. Although similar contemporary studies recognize these aspect of attention [46–50], only one study verified biological interpretability with an expert pathologist [51]. Further, no studies compare model performance to pathologist performance or non-expert performance, verify model-identified key instances with current practices, or recognize the potential for their model to discover novel imaging biomarkers. As evidenced by this study, the field of computational pathology may benefit from the proposed machine learning model's interpretability, as it provides a mechanism for the automatic discovery of novel image biomarkers as well as validation of current practices.

2. Methods

2.1. Study design

2.1.1. Ethics statement

All procedures were approved by Tufts University's Institutional Animal Care and Use Committee, and the Institutional Biosafety Committee. These experiments were approved under IACUC protocols: G2012-53; G2015-33; G2018-33. Biosafety Level 3 (BSL3) work was approved by under IBC registrations GRIA04; GRIA10; and GRIA17.

2.1.2. Mice and M.tb infection

Female DO mice (n = 452) and female C57BL/6J inbred mice (n = 30) from The Jackson Laboratory (Bar Harbor, ME) and were housed with sterile caging, bedding, food, and water in the biosafety level 3 (BSL3) facility at the New England Regional Biosafety Laboratory (Tufts University, Cummings School of Veterinary Medicine, North Grafton, MA). At 8-10-weeks old, mice were infected with low (~100 bacilli, n = 176) or very low (~20 bacilli, n = 256) dose of *M.tb* strain Erdman bacilli using a CH Technologies nose-only system, as we have described [23]. Mice were randomly assigned to cages prior to infection. Mice were monitored for health daily, weighed thrice weekly, and euthanized when signs of morbidity due to pulmonary TB developed (i.e. loss of body condition, respiratory distress). All mice were confirmed to be infected with *M.tb*. Age and gender matched non-infected control DO mice (n = 40) were identically housed, monitored, and euthanized at experimental end points. We selected a DO sample size to provide sufficient power to detect a quantitative trait locus that accounts for 10% of the phenotypic variance with 80% power at an alpha of 0.5 based on power simulations presented in [13]. We also provided enough mice to have at least 140 mice in each of the susceptibility classes.

2.1.3. Genome construction of DO mice

DO haplotypes are constructed using allele calls from GigaMUGA mouse genotyping array performed by Neogen (Lincoln, NB) and a hidden Markov model (HMM) implemented in R software as described [9].

2.1.4. Diagnostic categories for M.tb infected mice

The "supersusceptible" and "not-supersusceptible" ground truth labels reflect clinical outcomes that occurred during experimental *M. tb* infection. Supersusceptible DO mice developed morbidity and mortality within 8 weeks of *M.tb* infection. Mice that survived 8 weeks without morbidity or mortality were "not-supersusceptible." These two phenotypes are robust and reproducible, as we have previously observed [21,22]. Mice were excluded if they were euthanized due to non-TB disease based on clinical examination and necropsy findings. The result total number of supersusceptible mice was 148, and the total number of not-supersusceptible mice was 266.

2.1.5. Slide preparation and digital images

After euthanasia, lungs from each mouse were inflated and fixed in 10% neutral buffered formalin, processed and embedded in paraffin, sectioned at 5μm and stained with hematoxylin & eosin (H&E) at Tufts University, Cummings School of Veterinary Medicine, Core Histology Laboratory (North Grafton, MA). H&E stained glass slides were magnified 400 times and digitally scanned by Aperio ScanScope at 0.23 microns per pixel (The Ohio State University's Comparative Pathology and Mouse Phenotyping Shared Resources Core Facility, Columbus, OH). The median image size was 153,384 × 82,575 pixels. A flow diagram for the study design can be found in Fig. 1a.

2.2. Model description

2.2.1. Multiple Instance Learning (MIL)

MIL is a machine learning method where weak labels are assigned to collections (called bags) rather than individual examples (called instances) like in conventional machine learning. Classification by MIL is performed at the bag level and not the single instance level like in supervised learning [52]. The underlying assumption is that one class ("positive") shares features with a second class ("negative") *and* possesses features unique to itself. Here, the "bags" are the whole slides with one of two possible ground truth class labels – supersusceptible and not-supersusceptible. The "instances" are the unannotated, small images sampled from H&E-stained lung sections. The MIL paradigm applies because lungs from *M.tb*-infected supersusceptible DO mice share microscopic features with lungs of mice not-supersusceptible (e.g., regions of normal lung tissue, lymphocytes, plasma cells, macrophages) *and* also contain unique features (e.g. large necrotizing granulomas infiltrated by many neutrophils).

2.2.2. Attention-based pooling

The attention-based pooling mechanism [45] automatically learns to dynamically weight embedded instances into a bag-level feature vector that is subsequently classified. For example, if we have a single mouse, we take images sampled (instances) from its digital H&E tissue section (bag) and extract features from each, forming instance embeddings. A weight is automatically computed for each embedded instance through the attention-based pooling mechanism, then a weighted sum combines them into a single, bag-level instance, corresponding slide-level embedding. Classification is then performed on this bag-level embedding.

$$z = \sum_{k=1}^{K} a_k h_k$$

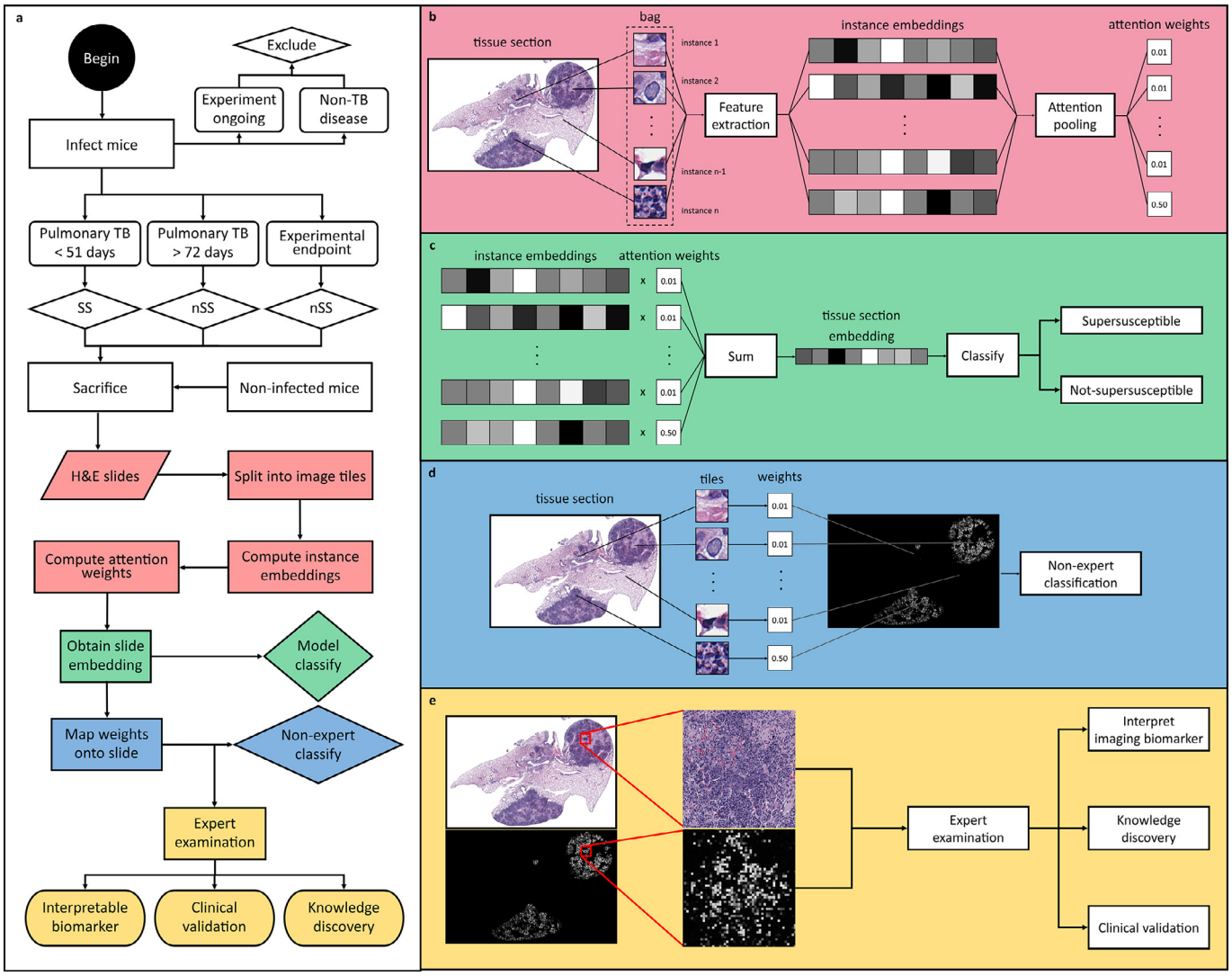

Fig. 1. The overall proposed methodology. a) The overall flowchart and proposed methodology. b) A bag of instances are created from a tissue section by sampling some number of image tiles. Each instance is individually subjected to feature extraction, resulting in instance features. Attention-based pooling computes attention weights (relative importance) of each of these instance vectors. c) Computed instance weights scale their respective instance features and are summed to a bag-level feature vector, which is classified as supersusceptible or not-supersusceptible. d) Attention weights are mapped back onto the tissue section image to bring attention to which areas of the tissue are being used for classification. e) An expert examines model-identified regions and identifies what the model attends to (i.e. "sees") leading to interpretable imaging biomarkers and potential knowledge discovery as well as validation of clinical practices.

$$a_k = \frac{exp\{w^T tanh(Vh_k^T)\}}{\sum_{j=1}^{K} exp\{w^T tanh(Vh_j^T)\}}$$

Our attention mechanism implementation consists of a simple two-layer fully connected network which passes each instance embedding (h_k) though one layer of the network (V), applies a tanh activation function to the result, then passes the activation though the second layer (w^T), which maps the vector into a single value, its attention weight (a_k). The weighted sum of each embedded instance and its attention weight yields a bag-level instance (z). The parameters (V,w) for this for this two-layer neural network are automatically learned through training of the model.

In addition to performing better than instance-based and embedding-based max and mean pooling approaches, the resulting instance weights allow the model to be *interpretable* in that the relative magnitudes of instance weights directly correspond to the instance's relative contribution to the overall classification of the bag [45]. Practically speaking, this means the model automatically identifies regions of H&E slides (i.e. an imaging biomarker) that contributes to its overall decision to classify a mouse as supersusceptible.

2.2.3. Model implementation

We modified our original model [45] because baseline model accuracy did not exceed pathologist performance. First, we decreased the second filter size from 3 × 3 to 1 × 1 (*Layer #3* in Supplemental Table 2) to improve the model's capacity to learn relationships across feature maps of the previous layer while reducing the number of parameters. Second, we increased the initial filter size from 4 × 4 to 7 × 7 (*Layer #1* in Supplemental Table 2) to eliminate the shifting caused by an even-dimensioned filter in the original model. This also increased the field of view for the initial set of filters. Further, this filter increased the perceptive field to a size that approximates the original two 3 × 3 convolutions. Overall, the 1 × 1 convolutions serve solely to learn and extract the relationship between feature maps while 7 × 7 convolution exploits the spatial relationships in individual feature maps. Third, we increased the corresponding number of filter maps in the first layer from 36 to 48 (*Layer #1* in Supplemental

Table 2) to offset the increase in the receptive field size when changing the filter size from 4 × 4 to 7 × 7 and to increase the feature set extracted by the network. Finally, we added an additional convolutional layer with ten 3 × 3 filters to refine and reduce the number of feature maps (*Layer #5* in Supplemental Table 2). Overall, our architecture requires fewer parameters (389,340) than the original (1,231,574) without increasing training time (ours: 14598±205s, original: 13216±2250s, p=0.19 using a paired t-test) across six folds and exhibits higher performance. A technical summary of the baseline model and the proposed model are in Supplemental Table 2.

Models were optimized using Adam with β_1 = 0.9 and β_2 = 0.999, the learning rate of 0.0001, weight decay of 0.0005, and over 100 or 200 epochs. Due to class imbalances, our cross-validation approach was carried by randomly sampling 30 tissue sections from both supersusceptible (~10%) and not-supersusceptible (~20%) categories for the validation set in each fold and including the rest of the cases for the training of the model, known as Monte Carlo cross-validation [53]. To account for imbalanced training sets, the training procedure was modified to randomly select supersusceptible mice or not-supersusceptible mice with equal likelihood and then to randomly select a mouse from the selected category during every training iteration. Negative log-likelihood was used as a cost function in our implementation.

2.2.4. Model summary

A bag is created from a tissue section by sampling some number of image crops (Fig. 1b). Each instance is subject to feature extraction using the implementation described in Supplemental Table 2, resulting in instance embeddings (Fig. 1b). Attention-based pooling computes attention weights (i.e. relative importance) of each of these instance vectors (Fig. 1b). Computed instance weights scale their respective instance embeddings and summed to a bag-level embedding (Fig. 1c). This embedding is classified as supersusceptible or not-supersusceptible (Fig. 1c) with a cutoff of 0.5. Attention weights for each image crop can then mapped back onto the tissue section image to bring attention to which areas of the tissue are being used for classification (Fig. 1d). The model was implemented in Python using the PyTorch library.

2.3. Analysis

2.3.1. Statistical methods

Model performance was evaluated using overall accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of the ten-fold cross-validation described above. 95% confidence intervals for each statistic above were computed using bootstrapped samples of predictions (equal to the number of observations) with replacement (n = 1000). 97.5th and 2.5th percentiles were taken as bounds for confidence intervals. Pathologist and non-expert performance were similarly evaluated using accuracy, sensitivity, specificity, positive predictive value, and negative predictive value with confidence intervals. The proposed machine learning model was compared to the baseline using a paired t-test of the distributions of accuracy of each fold. Each procedure was carried out in MATLAB. Multivariate analyses used ANOVAs followed by Kruskal-Wallis post-tests in GraphPad Prism v8 for the comparison of the quantified imaging biomarker.

3. Results

3.1. Lung granuloma patterns of M.tb-infected DO mice resemble humans

We observed many lung granuloma patterns in *M.tb*-infected DO mice (Fig. 2), mimicking the heterogeneity of granulomas in human pulmonary TB patients [54]. The lungs of supersusceptible DO mice, who 1) developed morbidity and mortality due to necropsy and 2) microbiologically-confirmed pulmonary TB within 8 weeks of infection, contained large coalescing regions of necrotic cellular, nuclear, and proteinaceous debris; neutrophils and macrophages; and tissue damage due to alveolar septal necrosis and capillary thrombosis (Fig. 2a–f). Lymphocytes and plasma cells were present but not prominent. The lungs of DO mice who survived more than 8 weeks without morbidity/mortality (i.e. not-supersusceptible) showed different granuloma patterns. Granulomas were not necrotic and instead contained many perivascular and peribronchiolar lymphocytes, plasma cells, and foamy macrophages (Fig. 2g–l) similar to most inbred mouse strains [55,56]. Interestingly, a small fraction (3–4%) of those DO mice developed histiocytic pneumonia, and lung cavities containing neutrophil and macrophage cellular debris surrounded by peripheral fibrosis (Fig. 2m–o). Chronic, discrete, organizing granulomas were occasionally observed (Fig. 2p–r). The lung patterns indicate that granuloma and/or lung tissue necrosis and neutrophilic influx are disease pathways of supersusceptibility, reflecting genetically controlled inflammatory responses to *M.tb*. These findings align with our prior work which showed that neutrophil chemokines (e.g. CXCL1, CXCL2, and CXCL5) and other innate inflammatory cytokines (e.g. Tumor Necrosis Factor) were significantly increased in the lungs of supersusceptible DO mice [23].

3.2. Model performance increases as the number instances included from each slide increases

Our first set of experiments explored how the baseline model would perform using different tile size and number of instances combinations (Supplemental Table 1). For each number of instances and tile size combination (Supplemental Table 1), a distinct dataset was generated. Each dataset consisted of a bag for each tissue section labeled with its associated diagnostic category (supersusceptible vs. not-supersusceptible). This resulted in 18 distinct datasets. Next, in order to determine optimal tile size and bag size configuration, a ten-fold cross-validation of the baseline model [45] was performed using the same set of training parameters (Supplemental Fig. 2). Overall, these experiments show two trends. Comparing configurations in the same rows (in which bag size remains the same), both training losses and errors quickly converge, albeit more slowly for smaller tile sizes, validation error generally doesn't converge, and validation loss begins to rise after about the same epoch. This implies that tile size does not affect the performance of the model. Comparing configurations within the same column (where tile size remains the same), as the number of tiles increases, convergence is slower but overall validation loss and error diverge less. This implies that the number of tiles is an important factor for validation convergence. Intuitively, this makes sense, as 1) the filter sizes in the CNN are much smaller than any tile size, so the resulting feature maps change little with change in tile size, and 2) a greater number of tiles represents more spread across the slide.

3.3. Proposed machine learning model significantly outperforms baseline model in identifying supersusceptibility

Using these optimized parameters (32 × 32 pixel tiles, mean/std bag size of 5000/1000), the proposed machine learning model (Supplemental Table 2) was subject to a ten-fold cross-validation and compared against the baseline model (Supplemental Fig. 3). The proposed machine learning model tends to perform much accurately (converging around 8% error) and converges at a faster rate compared to approximately the baseline model (approximately 16% error). The resulting cross-validation accuracy is 89.00 ± 2.96% (95% CI [86.83,91.00]) for the baseline model and 91.50 ± 4.68% (95% CI [89.83,93.33]) for the proposed machine learning model (paired Student's t-test, P = 0.04). Sensitivity, specificity, positive predictive value, and negative predicate value for the baseline model are 87.67, 85.00, 85.39, and 87.33 with

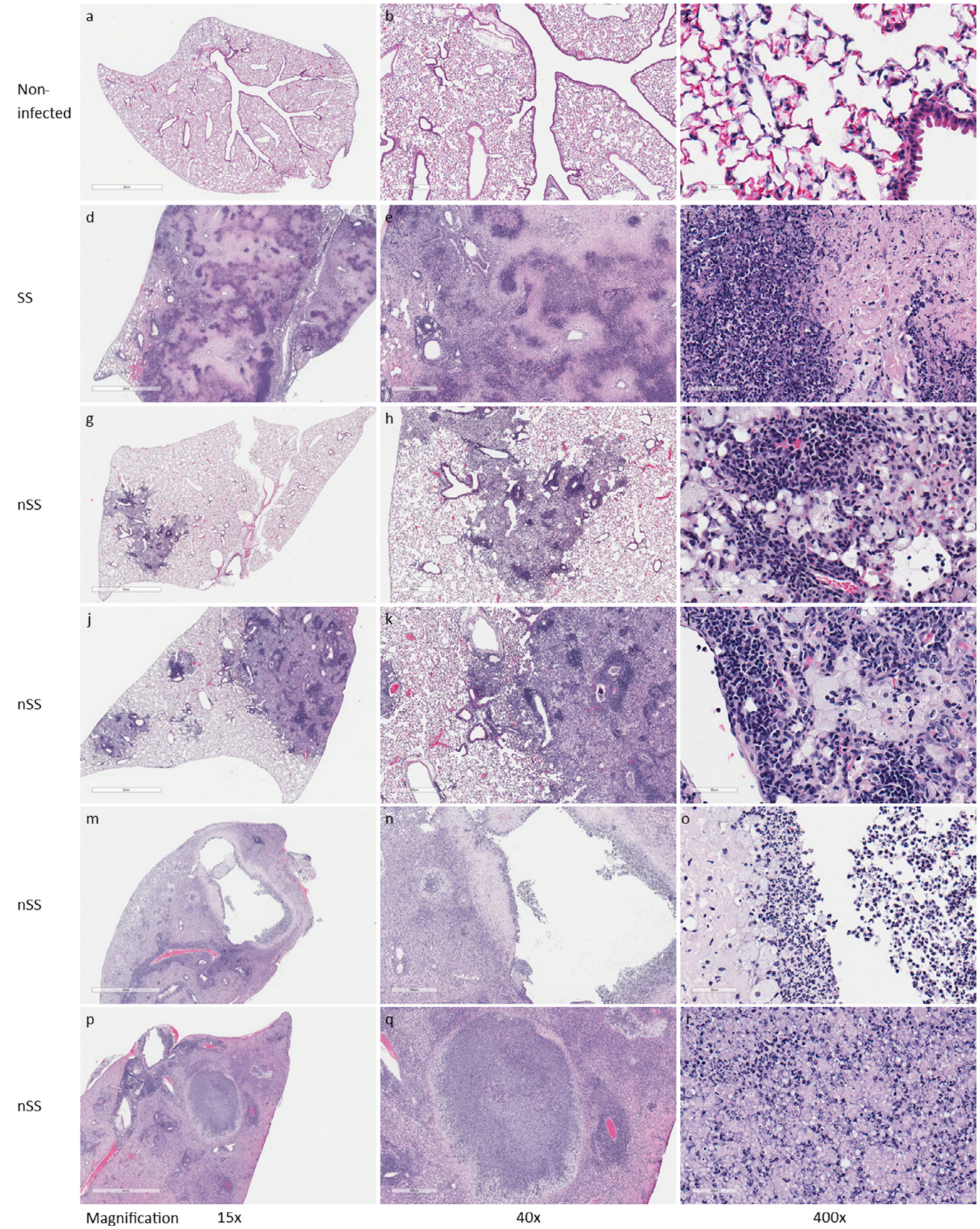

Fig. 2. Examples of lung sections from M.tb-infected DO mice. 8-10-week-old female DO (J:DO) and C57BL/6J mice were infected with 25±7 of aerosolized M.tb bacilli and euthanized when signs of morbidity developed or by day 150 after infection (whichever came first). Lung lobes were fixed in formalin, paraffin-embedded, sectioned at 5μm, and stained with hematoxylin and eosin. Lungs from normal, non-infected mice are shown in panels a, b, c. Lungs from supersusceptible mice with granuloma necrosis with neutrophil influx and pyknotic nuclear debris are shown in panels d, e, f. Lungs from not-supersusceptible mice are a spectrum from mild (g, h, i) to moderate (j, k, l) lymphohistiocytic pneumonia to severe pneumonia with cavitation and fibrosis (m, n, o), and necrosuppurative granulomas (p, q, r). Images are magnified 15x, 40x, and 400x normal from left to right.

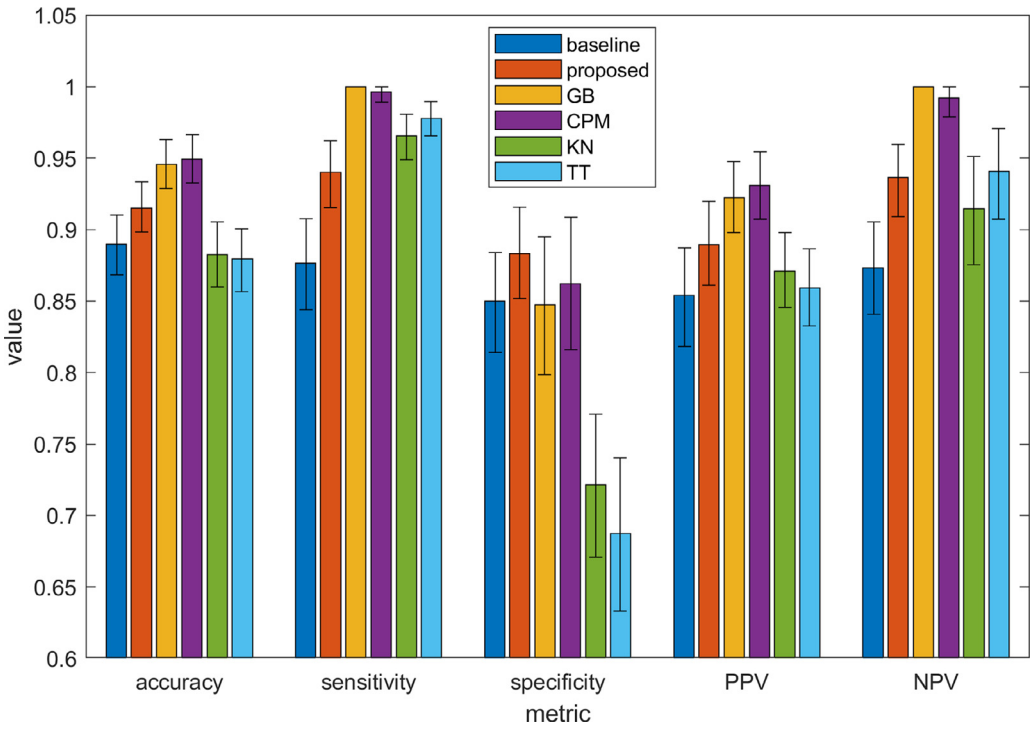

Fig. 3. Pathologists, non-expert, and model performance identifying supersusceptibility. Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) for the baseline model (baseline) using H&E, proposed model (proposed) using H&E, expert pathologist (GB) using H&E, resident pathologist (CPM) using H&E, non-expert (KN) using heatmaps, and non-expert (TT) using heatmaps for identifying supersusceptible mice.

95% CI [84.42,90.76], [81.42,88.38], [81.80,88.74], and [84.06,90.53], respectively, and for the proposed machine learning model 94.00, 88.33, 88.96, and 93.64 with 95% CI [91.54,96.21], [85.17,91.57], [86.12,91.97], and [90.92,95.95], respectively

3.4. Proposed machine learning model approaches pathologist performance to diagnose supersusceptibility

We compared the model's performance to diagnose supersusceptibility using lung tissue sections against one expert board-certified veterinary pathologist (Gillian Beamer) with over ten years of experience studying lungs of mice experimentally infected with *M.tb* and one board-eligible third-year veterinary pathology resident (Cesar Piedra-Mora) with no experience in mouse TB model. The pathologists reviewed H&E stained lung sections blinded. The experienced pathologist (GB) directly evaluated all sections with no additional training, while the junior pathologist (CPM) studied 10 randomly selected images from each class (supersusceptible and not-supersusceptible) prior to evaluating all available images 1 week later. The expert pathologist's accuracy on H&E stained lung tissue sections from 424 mice was 94.58% (95% CI [92.86,96.31]). Their sensitivity, specificity, positive predictive value, and negative predictive value were 100.00, 84.72, 92.25, and 100.00, respectively, with 95% CI [100.00, 100.00], [79.87,89.51], [89.78,94.74], and [100.00,100.00], respectively. The junior pathologist's accuracy was 94.95% (95% CI [93.27,96.63]). Their sensitivity, specificity, positive predictive value, and negative predictive value were 99.63, 86.21, 93.10, and 99.21 with 95% CI [98.94,100.00], [81.61,90.85], [90.71, 95.45], and [97.90,100.00]. By cross-validation, the proposed machine learning model performed similarly to experienced pathologists (Fig. 3).

3.5. Automatically identifies pyknotic nuclei and nuclear debris as imaging biomarkers for supersusceptibility

Each slide was exhaustively divided into 32 × 32 image crops, and their attention weights were computed using proposed machine learning model. A corresponding image heatmap was generated in which image crops were replaced by their attention weights. Resulting examples can be seen in Figs. 4 and 5 and Supplemental Fig. 4. In order to ascertain interpretability, potential as automatically identified imaging biomarkers, and clinical validity, the expert pathologist (GB) examined resulting heatmaps alongside respective tissue section counterparts and identified what the model was attending to (Fig. 1e). She concluded that the primary feature being identified by the proposed machine learning model was pyknotic nuclei and nuclear debris (Fig. 5). One other board-certified veterinary pathologist (Famke Aeffner) corroborated these observations by judging the presence of necrosis from 30 images sampled model-identified regions (like Fig. 5a–d) and non-model-identified regions (like Fig. 5e–h). They agreed with the model's identification of necrosis 73.33% of the time.

3.6. Non-experts identify supersusceptible mice using imaging biomarkers with high accuracy

After generating heatmaps for each digital tissue section, two non-experts (Thomas Tavolara and Khalid Niazi) were trained on a subset of slides (30 supersusceptible and 30 not-supersusceptible class) and validated on the rest (364 slides). Their respective accuracies for identifying supersusceptible and not-supersusceptible mice were 88.25% (95% CI [85.99,90.55]) and 87.95% (95% CI [85.67,90.07]) (Fig. 3). Their respective sensitivities, specificities, positive predictive values, and negative predictive values were 96.55, 72.12, 87.11, and 91.46 with 95% CI [94.89,98.08], [67.07,77.08], [84.55,89.78], and [87.54,95.13], respectively, and 97.78, 68.75, 85.93, and 94.08 with 95% CI [96.57,98.96, 63.29,74.02, 83.24,88.66], and [90.73,97.08], respectively. Their agreement was high and is depicted in Table 2.

3.7. Automatic quantification of imaging biomarkers provides continuous data for statistical analyses

Pathologists cannot quantify cellular, nuclear, or stromal features within a tissue section with accuracy or reproducibility [57,58].

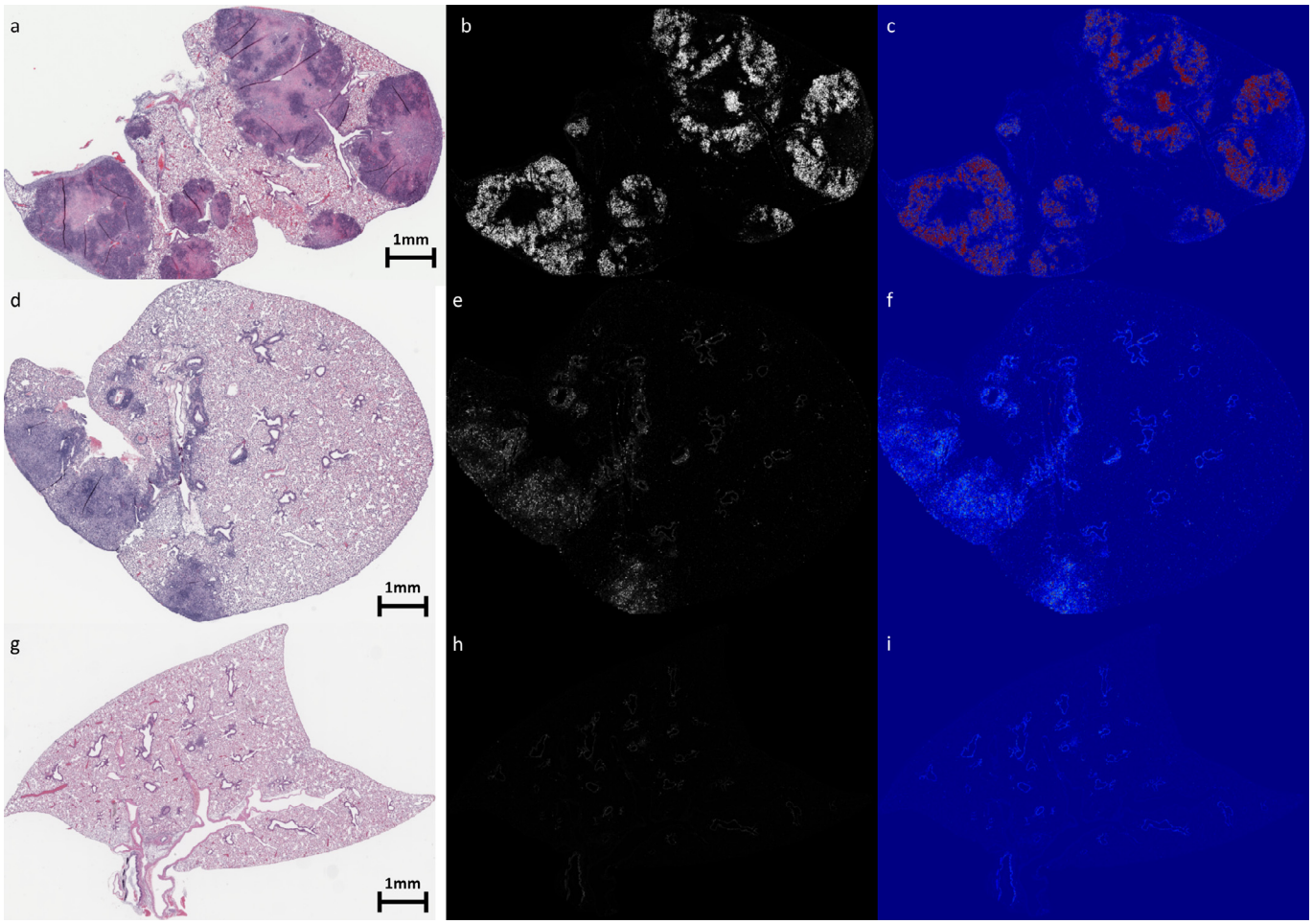

Fig. 4. Heatmap examples. The top set is a representative example of the supersusceptible class. The middle set is a representative example of the not-supersusceptible category. The bottom set is a representative example of non-infected mice. Heatmap images are contrast enhanced for visual purposes. Brighter (middle column) or redder (right column) regions correspond to areas of the images that correspond to supersusceptible features (i.e. image biomarkers). All images are taken at 1x magnification.

Generally, they attempt to judge extent of disease by relying on grades or scales that are estimates [59,60]. The automatically identified lung imaging biomarker was quantified (Supplementary Methods) for tissue sections in each mouse as a ratio of necrotic pyknotic affected regions to non-necrotic tissue (Fig. 6). The imaging biomarker was significantly increased in the lung sections of *M.tb*-infected supersusceptible DO mice (n = 148) than not-supersusceptible (n = 266) *M.tb*-infected DO mice, *M.tb*-infected C57BL/6J inbred mice (n = 10), and non-infected DO mice.

4. Discussion

Each DO mouse carries genome segments from 8 inbred founder strains, resulting in a mosaic across the chromosomes (Supplementary Fig. 1). The genetic diversity of DO mice is equivalent to humans [8], and this genetic diversity contributes to unique and wide phenotypes [20] that are not observed within or between common inbred strains [5,6,19,22–30]. Even our work with panels of founder and Collaborative Cross inbred strains [29] do not produce the response range in DO mice [23] and our unpublished data. We showed that supersusceptible DO mice develop pulmonary TB with large necrotizing granulomas and abundant *M.tb* bacilli, while those that survive longer develop non-necrotic lymphocyte-rich granulomas. Published elsewhere, we confirmed that DO mice are not immune deficient or highly susceptible to other mycobacteria, and they generate antigen-specific immunity against *M.tb* [16,23,24]. Together, these results support our conclusion that DO mice are a valuable animal model of TB. Since we can now rigorously quantify necrotic and pyknotic debris within lung granulomas using the method here, our ongoing and future work can focus identifying novel mechanisms of *M.tb*-induced granuloma necrosis. This is a great translational benefit of DO mice and of computational pathology because tissue and cellular features of histology images cannot be accurately quantified by pathologists in large image data sets.

Although deep learning in the context of computer vision has achieved many great feats, from human-level object detection to realistic synthetic image generation and self-driving vehicles, it has still generally been eluded by knowledge discovery, in which other fields have seen some successes [61,62]. The results of the proposed machine learning model suggest a path for such knowledge discovery in medical image analysis. Here, we did not use any manual annotations to guide the MIL. Instead, two outcomes (classes) were used solely – supersusceptible, defined as rapid pulmonary TB within 8 weeks of infection; and not-supersusceptible, defined as no clinical signs for at least the same duration. This identified an imaging biomarker in granulomas that 1) Automatically diagnoses supersusceptibility; 2) Serves as a tool (heatmap) for non-experts to diagnose supersusceptibility; 3) Was clinical interpretable by two board-certified veterinary pathologists as pyknotic debris; and 4) Quantifies pyknotic debris for statistical comparisons. Our results imply that computational approaches can produce and use the same imaging biomarkers that pathologists use, mimicking pattern recognition and decision-making of pathologists reading

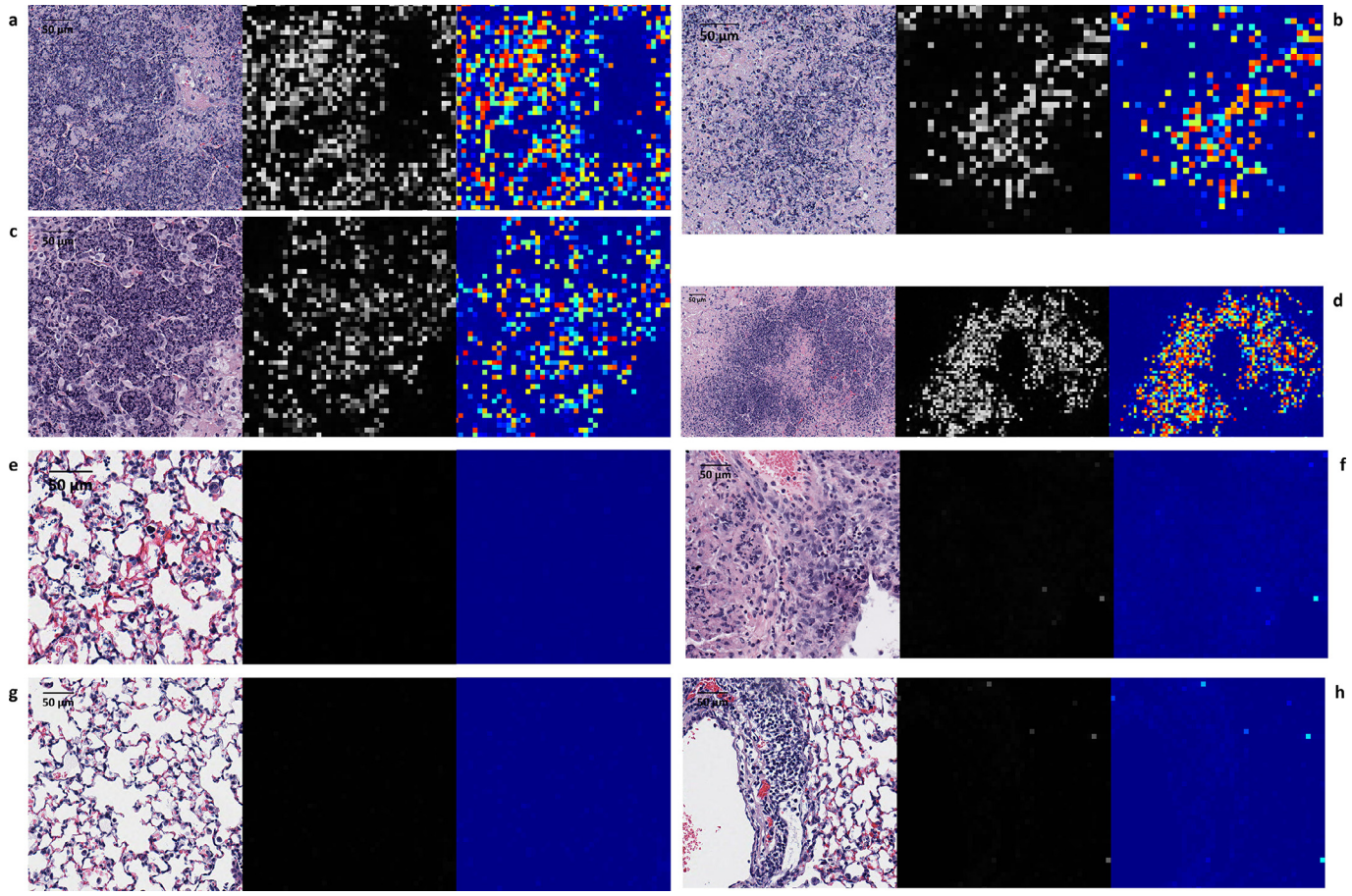

Fig. 5. Primarily nuclear debris is being identified. Heatmaps with their corresponding tissue region to the right (H&E) from eight supersusceptible mice. Brighter (middle column) or redder (right column) regions correspond to areas of the images that correspond to supersusceptible features (i.e. image biomarkers). Note that primarily nuclear debris is being identified and that healthy tissue is not being identified.

histology slides. Thus, our work sheds light on the "black box" of computational pathology and instills confidence that artificial intelligence can make accurate diagnoses using interpretable, biologically relevant histopathology imaging biomarkers.

We recognize three major shortcomings of conventional deep learning approaches [34,36–41,58,63,64] which the proposed MIL model overcomes. First, conventional deep learning models require *strong* labels [34]. Labels are classes or identities we assign to images. Strong labels can be thought of as any label that can be delineated in an image (a.k.a. image annotations). For example, in the context of medical image analysis, a strong label could be an outline or segmentation of an organ or tissue structure in MRI or CT, or it could be anatomical structures or cell types in histology. The proposed MIL model directly addresses this shortcoming of conventional approaches by requiring *weak* labels, labels assigned to whole image rather than its parts. Examples of weak labels include a diagnosis, status, tumor malignancy/benignness, cancer stage, or degree of neurodegeneration. In the context of the problem addressed here, the *weak* labels of supersusceptible and not-supersusceptible are utilized not only because they are available to us but also because tissue-level knowledge (i.e. *strong* labels) is not readily available – there are no tissue

Table 2
Agreement between two non-experts using imaging biomarkers.

	Non-expert 1 predicts not SS	Non-expert 1 predicts SS
Non-expert 2 predicts not SS	262	11
Non-expert 2 predicts SS	4	86

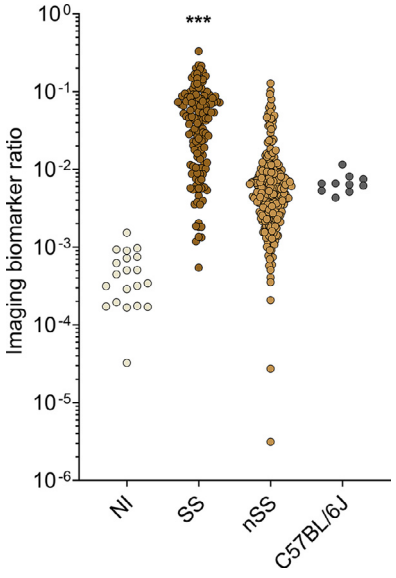

Fig. 6. Quantification of image biomarker for statistical analysis. The imaging biomarker was quantified as a ratio of necrotic/pyknotic affected regions to non-necrotic tissue and analyzed across multiple groups including: non-infected (NI) DO mice ($n = 20$) and M.tb-infected DO mice in supersusceptible ($n = 148$) and not-supersusceptible ($n = 266$) categories. The biomarker was also quantified in the lungs of M.tb-infected C57BL/6J inbred mice ($n = 10$). The lungs of supersusceptible mice contain significantly more of the imaging biomarker than M.tb-infected not-supersusceptible DO mice, C57BL/6J inbred mice, and non-infected DO mice. Each dot represents one individual mouse. Data was analyzed by ANOVA with multiple comparisons and Kruskal-Wallis posttest ($^{***}p < 0.001$).

areas or cell types which we can easily and confidently annotate as super-susceptible or not-supersusceptible. The distinction is key for the automated analysis of medical images, as acquisition of relevant annotations (i.e. domain knowledge) is time consuming and thus one of the barriers for artificial intelligence in medicine.

Second, conventional deep learning models fall short, as they do not efficiently process digital tissue slides, which are in the order of gigapixels. TB detection studies that utilize chest x-ray images often rely on CNNs for fibrosis detection [36–41]. Chest x-ray image sizes are in the order of thousands by thousands of pixels (i.e. 1000×1000 to 5000×5000), and as is typical in CNN applications, are resized to small images (typically 200×200 to 300×300) 1) for the purpose of fitting the fixed-input size for their CNN model and 2) because of lack of memory on current GPUs. Though not explicitly justified in the literature, the underlying assumption in resizing images is that smaller details (which are lost from resizing) aren't necessary for CNN models. Unlike x-rays, digital pathology slides are in the order of billions of pixels. To resize an entire slide to the fixed-input size of a deep learning model would result in loss of nearly all cell and anatomical structure information. Thus, we cannot simply resize the input image. Typical solutions in digital pathology include extracting high-power fields (akin to fields of view a pathologist might analyze under the microscope) or extracting tiles (small images cropped from the slide). In the former, high power fields are annotated by a pathologist, thereby returning to the problems of strong labeling discussed earlier. In the latter, individual tiles are assigned labels based on their anatomical content. However, processing hundreds of thousands of image tiles per slide is inefficient in addition to requiring strong labeling. Moreover, it is not explicitly clear how model-predicted cell and tissue level labels relate to weaker clinical labels nor how multiple tile-level predictions should be aggregated into clinical labels. The proposed MIL model resolves the magnitude of the problem (i.e. the large image size) by sampling digitized tissue sections (without resizing) for small 32×32 image tiles. It also resolves the question of how to aggregate said tiles through an automatically learned aggregation mechanism – namely, attention pooling.

Third, conventional deep learning models are generally not interpretable. Lack of interpretability is an oft criticized aspect of CNN-based deep learning in image analysis [58,63,64]. Though CNN models have been shown to achieve near human-level performance on several computer vision tasks in specific domains, they generally lack the ability to explain how they decide. Thus, such approaches are often dubbed 'black-boxes.' This black-box nature is particularly troubling in biomedical research and medicine, as scientists and clinicians need to know how and why a model makes decisions in order to understand some underlying phenomenon (i.e. knowledge discovery) and to make informed clinical decisions. Given this crucial shortcoming of CNN-based deep learning, there exist several tools to give some insight into what they "see." Class activation mapping (CAM) [42] was one of the first of these tools. This method allowed CNN models to highlight which parts of an input image contribute to its overall classification. Grad-CAM [43] immediately improved on CAM by allowing for more complex CNN models. There was further improvement in a method dubbed Grad-CAM++ [44]. Several studies examining CNNs as potential clinical tools for TB utilized these methods to visualize which parts of chest X-rays were being utilized for their model's decision [36–39]. However, in digital pathology, these methods cannot be applied, as they require entire images as input to highlight decision-supporting regions. Digital slides are too large to be processed in such a manner and cannot be resized for reasons discussed. Further, these methods focus largely on strong labels in multi-class problems and thus cannot apply to the current problem. The proposed MIL model resolves these issues of interpretability through the automatically learned attention pooling mechanism [45]. Automatically identified instances give insight into what the proposed machine learning model is "seeing."

In the context of medical image analysis, this attention pooling provides a useful mechanism that does not require manual annotation of microscopic features. Here, it automatically discovered an imaging feature ("biomarker") diagnostic for supersusceptibility – namely, pyknotic nuclei and nuclear debris. Despite the proposed machine learning model not being aware, granuloma necrosis is used by pathologists to diagnose pulmonary TB in human patients and is a desirable feature of experimental TB in animal models. Thus, our automatic method mimics how pathologists examine biopsy, autopsy, and experimental tissues to make diagnostic decisions, providing validation and clinical relevance. Importantly, the proposed machine learning model did not identify healthy lung tissue. This affirms the MIL assumption that positive bags may contain instances similar to those found in negative bags. In this problem, those instances are healthy lung tissue (Fig. 5e–h). Healthy tissue is present in both supersusceptible and not-supersusceptible *M.tb*-infected DO mice. It therefore should not be a discriminatory feature identified by the proposed machine learning model. And indeed, this is the case. It is obvious from the heatmap that areas of healthy lung tissue and never attended to (Figs. 4 and 5e–h).

As a result of automatically identifying imaging biomarkers, the proposed machine learning model was also able to generate heatmaps corresponding to presence and intensity of features indicative of supersusceptibility. This resulted in simplified images which non-experts used to classify supersusceptibility or not with greater accuracy than an expert using H&E. This is understandable, as pathologists are trained for years to differentiate among thousands of different tissues, cell types, and other features in more than one type of stain. Comparatively, the proposed machine learning model learns only to focus a few of these features, primarily pyknotic nuclei and nuclear debris. Additionally, it samples across the entire slide rather than a small, localized subset of regions (like pathologists). Non-experts benefit from heatmaps by being able to focus on a specific feature while being undistracted by irrelevant features. Both non-expert readers identified two key features of heatmaps in their training set – density and clustering. In general, supersusceptible heatmaps were identifiable by multiple large, dense clusters. Heatmaps from the not-supersusceptible class usually were either not very dense or did not form clusters (Fig. 4; Supplemental Fig. 4).

The proposed machine learning model has limitations. For example, it cannot identify imaging biomarkers of non-supersusceptibility, likely because the non-supersusceptible class is homogenous, containing few morphological subtypes and also non-infected DO mice. Future work will yield imaging biomarkers of intermediate and long-term resistance to *M.tb*. Currently, we are increasing our data set and investigating how to automatically discover important features of other classes of *M.tb*-infected DO mice, such susceptible and resistant DO mice (Table 1).

Overall, our results show that computational approaches can automatically identify and utilize the same imaging biomarkers that pathologists use, mimicking expert human pattern recognition and decision-making. Our work sheds light on the "black box" of computational pathology and instills confidence that artificial intelligence methods can make accurate diagnoses using interpretable histopathology features. In future studies, we intend to develop a multi-class version of this framework to identify and quantify unique granuloma features of highly resistant DO mice, and of vaccinated DO mice. Finally, we will explore how the proposed machine learning model performs in related scenarios and whether model-identified imaging biomarkers are interpretable quantifiable.

Contributors

TET contributed to data curation, formal analysis, methodology, software, validation, writing the original draft, and editing the manuscript. MKKN contributed to methodology, formal analysis, and

review and editing of the manuscript. MG contributed to data collection. CPM contributed to review and editing of the manuscript. DG contributed to data analysis and review and editing of the manuscript. GB contributed to data collection, data curation, formal analysis, conceptualization, funding acquisition, project administration, and review and editing of the manuscript. MNG contributed to conceptualization, formal analysis, funding acquisition, methodology, project administration, and review and editing of the manuscript.

Data sharing statement

All data and code utilized for development and validation is available at github.com/cialab/imaging_biomarker to anyone who wishes to utilize it for any purpose following publication. All raw imaging data will be made available upon request by contacting mniazi@wakehealth.edu immediately after publication.

Declaration of Competing Interest

Mr. Tavolara has nothing to disclose. Dr. Niazi has nothing to disclose. Ms. Ginese has nothing to disclose. Dr. Piedra-Mora has nothing to disclose. Dr. Gatti has nothing to disclose. Dr. Beamer has nothing to disclose. Dr. Gurcan has nothing to disclose.

Acknowledgements

We thank Aubrey Specht (Tufts Cummings School of Medicine) for her time and effort using the heatmaps to classify supersusceptible and not-supersusceptible mice. We thank Ms. Julie Tzipori, Mr. Curtis Rich, Mr. Donald Girouard, and Dr. Samuel Telford at the NE-RBL, Tufts University Cummings School of Veterinary Medicine, North Grafton, MA. We thank Ms. Frances Brown, Ms. Linda Wrijil, Ms. Sarah Ducat, and Ms. Gina Scarglia for the histology services at Tufts University's Cummings School of Veterinary Medicine. We also would like to thank board-certified veterinary pathologist and Dr. Famke Aeffner (Amgen Inc.) for corroborating Dr. Gillian Beamer's interpretation of the imaging biomarker. Written consent was for acknowledgement was received from each individual. Support was provided by NIH R21 AI115038; NIH R01 HL145411; and the American Lung Association Biomedical Research Grant RG-349504 (GB and MG). The Comparative Pathology & Mouse Phenotyping Shared Resource, Department of Veterinary Biosciences and the Comprehensive Cancer Center, The Ohio State University, Columbus, OH, supported in part by grant P30 CA016058 provided Aperio ScanScope slide scanning services. The funders had no role in study design, data collection, data analysis, interpretation, or writing of the report.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.103094.

References

[1] World Health Organization (WHO). Tuberculosis 2018 [Available from: https://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.
[2] Achkar JM, Jenny-Avital ER. Incipient and subclinical tuberculosis: defining early disease states in the context of host immune response. J Infect Dis 2011;204 (suppl_4):S1179–S86.
[3] Leong FJ, Dartois V, Dick T. A color atlas of comparative pathology of pulmonary tuberculosis. CRC Press; 2016.
[4] World Health Organization (WHO). Tuberculosis global facts. Available from: https://www.who.int/tb/publications/factsheet_global.pdf.
[5] Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nat Genet 2007;39(9):1100–7.
[6] Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 2011;43(7):648–55.
[7] Bogue MA, Churchill GA, Chesler EJ. Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. Mamm Genome 2015;26(9-10):511–20.
[8] Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. Mamm Genome 2012;23(9-10):713–8.
[9] French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, et al. Diversity outbred mice identify population-based exposure thresholds and genetic factors that influence benzene-induced genotoxicity. Environ Health Perspect 2015;123(3):237–45.
[10] Gatti DM, Svenson KL, Shabalin A, Wu LY, Valdar W, Simecek P, et al. Quantitative trait locus mapping methods for diversity outbred mice. G3 2014;4(9):1623–33.
[11] Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, et al. High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics 2012;190(2):437–47.
[12] Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 2004;36(11):1133–7.
[13] Threadgill DW, Churchill GA. Ten years of the Collaborative Cross. Genetics 2012;190(2):291–4.
[14] Chesler EJ, Gatti DM, Morgan AP, Strobel M, Trepanier L, Oberbeck D, et al. Diversity outbred mice at 21: maintaining allelic variation in the face of selection. G3 2016;6(12):3893–902.
[15] Churchill GA, Gatti DM, Munger SC, Svenson KL. The diversity outbred mouse population. Mamm Genome 2012;23(9−10):713–8.
[16] Kurtz SL, Rossi AP, Beamer GL, Gatti DM, Kramnik I, Elkins KL. The diversity outbred mouse population is an improved animal model of vaccination against tuberculosis that reflects heterogeneity of protection. mSphere. 2020;5(2).
[17] Hunter RL, Actor JK, Hwang S-A, Karev V, Jagannath C. Pathogenesis of post primary tuberculosis: immunity and hypersensitivity in the development of cavities. Ann Clin Lab Sci 2014;44(4):365–87.
[18] Bourbonnais JM, Sirithanakul K, Guzman JA. Fulminant miliary tuberculosis with adult respiratory distress syndrome undiagnosed until autopsy: a report of 2 cases and review of the literature. J Intens Care Med 2005;20(6):306–11.
[19] Major S, Turner J, Beamer G. Tuberculosis in CBA/J mice. Veterinary pathology; 2013.
[20] Rasmussen AL, Okumura A, Ferris MT, Green R, Feldmann F, Kelly SM, et al. Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. Science 2014;346(6212):987–91.
[21] Kus P, Gurcan MN, Beamer GJM. Automatic detection of granuloma necrosis in pulmonary tuberculosis using a two-phase algorithm: 2D-TB. 2019;7(12):661.
[22] Mouse models of human TB pathology: roles in the analysis of necrosis and the development of host-directed therapies. In: Kramnik I, Beamer G, editors. Seminars in immunopathology. Springer; 2016.
[23] Niazi MK, Dhulekar N, Schmidt D, Major S, Cooper R, Abeijon C, et al. Lung necrosis and neutrophils reflect common pathways of susceptibility to Mycobacterium tuberculosis in genetically diverse, immune-competent mice. Dis Model Mech 2015;8(9):1141–53.
[24] Harrison DE, Astle CM, Niazi MK, Major S, Beamer GL. Genetically diverse mice are novel and valuable models of age-associated susceptibility to Mycobacterium tuberculosis. Immun Ageing 2014;11(1):24.
[25] Harper J, Skerry C, Davis SL, Tasneen R, Weir M, Kramnik I, et al. Mouse model of necrotic tuberculosis granulomas develops hypoxic lesions. J Infect Dis 2012;205 (4):595–602.
[26] Lyadova IV, Tsiganov EN, Kapina MA, Shepelkova GS, Sosunov VV, Radaeva TV, et al. In mice, tuberculosis progression is associated with intensive inflammatory response and the accumulation of Gr-1^{dim} cells in the lungs. PLoS One 2010;5(5): e10469.
[27] Eruslanov EB, Lyadova IV, Kondratieva TK, Majorov KB, Scheglov IV, Orlova MO, et al. Neutrophil responses to *Mycobacterium tuberculosis* infection in genetically susceptible and resistant mice. Infect Immun 2005;73(3):1744–53.
[28] Nandi B, Behar SM. Regulation of neutrophils by interferon-γ limits lung inflammation during tuberculosis infection. J Exp Med 2011;208(11):2251–62.
[29] Smith C, Proulx M, Olive A, Laddy D, Mishra B, Moss C, et al. Tuberculosis susceptibility and vaccine protection are independently controlled by host genotype. mBio 2016;7(5):e01516.
[30] Beamer GL, Turner J. Murine models of susceptibility to tuberculosis. Arch Immunol Ther Exp 2005;53(6):469–83.
[31] Niazi MKK, Beamer G, Gurcan MN. An application of transfer learning to neutrophil cluster detection for tuberculosis: efficient implementation with nonmetric multidimensional scaling and sampling. SPIE; 2018.
[32] Niazi MKK, Beamer G, Gurcan MN, editors. A computational framework to detect normal and tuberculosis infected lung from H and E-stained whole slide images. Medical Imaging 2017: digital pathology. International Society for Optics and Photonics; 2017.
[33] Tavolara TE, Niazi MKK, Beamer G, Gurcan MN, editors. Segmentation of mycobacterium tuberculosis bacilli clusters from acid-fast stained lung biopsies: a deep learning approach. Medical Imaging 2020: digital pathology. International Society for Optics and Photonics; 2020.
[34] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[35] Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw Open 2019;2(11) e1914645-e.
[36] Hwang S, Kim H-E, Jeong J, Kim H-J, editors. A novel approach for tuberculosis screening based on deep convolutional neural networks. Medical imaging 2016: computer-aided diagnosis. International Society for Optics and Photonics; 2016.

[37] Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574–82.
[38] Becker A, Blüthgen C, Sekaggya-Wiltshire C, Castelnuovo B, Kambugu A, Fehr J, et al. Detection of tuberculosis patterns in digital photographs of chest X-ray images using Deep Learning: feasibility study. Int J Tuberand Lung Dis 2018;22(3):328–35.
[39] Application of a Convolutional Neural Network using transfer learning for tuberculosis detection. In: Ahsan M, Gomes R, Denton A, editors. 2019 IEEE International Conference on Electro Information Technology (EIT). IEEE; 2019.
[40] Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2018;69(5):739–47.
[41] Lopes U, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Comput Biol Med 2017;89:135–43.
[42] Learning deep features for discriminative localization. In: Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, editors. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
[43] Grad-cam: visual explanations from deep networks via gradient-based localization. In: Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Proceedings of the IEEE international conference on computer vision; 2017.
[44] Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN, editors. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2018.
[45] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. arXiv preprint arXiv:180204712. 2018.
[46] Multiple instance learning for heterogeneous images: training a cnn for histopathology. In: Couture HD, Marron JS, Perou CM, Troester MA, Niethammer M, editors. International conference on medical image computing and computer-assisted intervention. Springer; 2018.
[47] Li J, Li W, Gertych A, Knudsen BS, Speier W, Arnold CW. An attention-based multi-resolution model for prostate whole slide imageclassification and localization. arXiv preprint arXiv:190513208. 2019.
[48] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8):1301–9.
[49] Distill-to-label: weakly supervised instance labeling using knowledge distillation. In: Thiagarajan JJ, Kashyap S, Karagyris A, editors. 2019 18th IEEE international conference on machine learning and applications (ICMLA). IEEE; 2019.
[50] Mercan C, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans Med Imaging 2017;37(1):316–25.
[51] Paschali M, Naeem MF, Simson W, Steiger K, Mollenhauer M, Navab N. Deep learning under the microscope: improving the interpretability of medical imaging neural networks. arXiv preprint arXiv:190403127. 2019.
[52] Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit 2018;77:329–53.
[53] Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemom Intell Lab Syst 2001;56(1):1–11.
[54] Leong FJW-M, Eum S, Via LE, Barry 3rd CE. Pathology of tuberculosis in the human lung. In: Leong FJ, Dartois V, Dick T, editors. A color atlas of comparative pathology of pulmonary tuberculosis. New York: CRC Press; 2011. p. 53–81.
[55] Niazi MK, Beamer G, Gurcan MN. Detecting and characterizing cellular responses to *Mycobacterium tuberculosis* from histology slides. Cytom Part A 2014;85(2):151–61.
[56] Vesosky B, Rottinghaus EK, Stromberg P, Turner J, Beamer G. CCL5 participates in early protection against *Mycobacterium tuberculosis*. J Leukoc Biol 2010;87(6):1153–65.
[57] Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener BJIribe. Histopathological image analysis: a review. 2009;2:147-71.
[58] Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019;20(5):e253–e61.
[59] Kim HB, Zhang Q, Sun X, Beamer G, Wang Y, Tzipori S. Beneficial effect of oral tigecycline treatment on Clostridium difficile infection in gnotobiotic piglets. Antimicrob Agents Chemother 2014;58(12):7560–4.
[60] Lee S, Ginese M, Beamer G, Danz HR, Girouard DJ, Chapman-Bonofiglio SP, et al. Therapeutic efficacy of bumped kinase inhibitor 1369 in a pig model of acute diarrhea caused by Cryptosporidium hominis. Antimicrob Agents Chemother 2018;62(7).
[61] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23(6):1241–50.
[62] Lan K, Wang D-t, Fong S, Liu L-s, Wong KK, Dey N. A survey of data mining and deep learning in bioinformatics. J Med Syst 2018;42(8):139.
[63] Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA 2018;320(11):1107–8.
[64] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236–46.