EBioMedicine 67 (2021) 103388

Contents lists available at ScienceDirect

EBioMedicine

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

Research paper

Deep learning predicts gene expression as an intermediate data modality to identify susceptibility patterns in *Mycobacterium tuberculosis* infected Diversity Outbred mice

Thomas E. Tavolara[a], M.K.K. Niazi[a,*], Adam C. Gower[b], Melanie Ginese[c], Gillian Beamer[c], 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 Medicine, Boston University School of Medicine, 72 E. Concord St Evans Building, Boston, MA 02118, United States
[c] Department of Infectious Disease and Global Health, Tufts University Cummings School of Veterinary Medicine, 200 Westboro Rd., North Grafton, MA 01536, United States

ARTICLE INFO

Article History:
Received 1 February 2021
Revised 22 April 2021
Accepted 23 April 2021
Available online xxx

Keywords:
Tuberculosis
Diversity Outbred mice
Gene expression
Deep learning
Histopathology

ABSTRACT

Background: Machine learning sustains successful application to many diagnostic and prognostic problems in computational histopathology. Yet, few efforts have been made to model gene expression from histopathology. This study proposes a methodology which predicts selected gene expression values (microarray) from haematoxylin and eosin whole-slide images as an intermediate data modality to identify fulminant-like pulmonary tuberculosis ('supersusceptible') in an experimentally infected cohort of Diversity Outbred mice (n=77).

Methods: Gradient-boosted trees were utilized as a novel feature selector to identify gene transcripts predictive of fulminant-like pulmonary tuberculosis. A novel attention-based multiple instance learning model for regression was used to predict selected genes' expression from whole-slide images. Gene expression predictions were shown to be sufficiently replicated to identify supersusceptible mice using gradient-boosted trees trained on ground truth gene expression data.

Findings: The model was accurate, showing high positive correlations with ground truth gene expression on both cross-validation (n = 77, $0.63 \leq \rho \leq 0.84$) and external testing sets (n = 33, $0.65 \leq \rho \leq 0.84$). The sensitivity and specificity for gene expression predictions to identify supersusceptible mice (n=77) were 0.88 and 0.95, respectively, and for an external set of mice (n=33) 0.88 and 0.93, respectively.

Implications: Our methodology maps histopathology to gene expression with sufficient accuracy to predict a clinical outcome. The proposed methodology exemplifies a computational template for gene expression panels, in which relatively inexpensive and widely available tissue histopathology may be mapped to specific genes' expression to serve as a diagnostic or prognostic tool.

Funding: National Institutes of Health and American Lung Association.

1. Introduction

Microscopic examination of histopathological tissue sections plays a central role in analysis, diagnosis and prognosis of biological tissues. Increasingly, whole-slide imaging of tissues, along with fast networks, data transfer and inexpensive storage has enabled the curation of large databases of digitized tissue sections [1]. Furthermore, rapid advances in deep learning methods have enabled scientists to develop automated histopathological analysis methods on whole-slide images (WSIs), ranging from primitives such as nuclei detection [2] and mitosis detection [3] to more advanced applications such as tumour grading [4]. By and large, these developments and those similar provide substantial evidence for the future of deep learning as an essential tool for clinical and biomedical fields.

While much effort continues to perpetuate the successful application of deep learning models to digital pathology, relatively few efforts have been made to connect histopathology to molecular markers such as gene mutations, gene transcripts and proteins. Recent studies have shown that deep learning can identify and localize areas of tissue correlated with specific mutations in the breast [5–7], lung [8], and liver [9] cancers; predict microsatellite instability in colorectal [10] and gastrointestinal tumours [11]; and predict tumour mutational burden in lung [12] and liver [13] cancers. The

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

https://doi.org/10.1016/j.ebiom.2021.103388

Research in Context

Evidence before this study

We performed literature searches and publications in English without date restrictions. We searched PubMed for "tuberculosis AND (mouse model) AND (gene expression) AND (susceptibility)" on January 31st, 2021. This retrieved 80 results – 77 were primary research articles and 3 were review articles. 69 of the primary research articles utilized inbred or gene-deleted mice whereas 8 utilized humanized mice or inbred mice with transgenes. No primary publications utilized Diversity Outbred mice. No primary research articles applied artificial intelligence or machine learning to any aspect of their analysis. We searched PubMed for the term "("deep learning") AND (histology OR pathology OR histopathology) and ("gene expression")" on January 31st, 2021. This retrieved 61 primary research articles and 1 review article. 45 of the primary research articles involved cancer. Four involved optimal drug prediction using gene expression. None involved tuberculosis. 26 utilized gene expression for some classification problem. Of these, one study utilized multiple instance learning, and four utilized gene expression for multi-modal analysis for some clinical outcome. Three regressed to gene expression. The same three predicted gene expression from haematoxylin & eosin-stained biopsies. However, none of these studies linked predicted gene expression to some diagnostic outcome.

Added value of this study

Attention-based deep learning can be applied to haematoxylin and eosin-stained tissues to predict expression of genes. Furthermore, when applied to predict genes which accurately discriminate disease outcomes of tuberculosis, the model is sufficiently accurate to replicate relevant gene expression values such that disease outcomes which are easily discriminated by microarray-based gene expression continue to be easily discriminated by the approximations made by the model.

Implications of all the available evidence

Our methodology can be used to map histopathology to gene expression with sufficient accuracy to predict a clinical outcome. The proposed methodology exemplifies a computational template for gene expression panels, in which relatively inexpensive tissue histopathology may be mapped to specific genes' expression to serve as a diagnostic or prognostic tool.

interest in predicting such signatures of disease stems from the downstream effects on phenotypes – particularly changes in gene expression – which in turn drives research towards particular targeted therapies [14] and informs clinical decisions [15,16]. Yet, save for one very recent study [17], no effort has been made to predict gene expression profiles from histopathology. Such a tool would prove invaluable, as current utilization of transcriptome analysis in the clinic is limited by cost, time, and standardization [18–20].

Our ongoing work focuses on utilization of histopathological and molecular data to identify biomarkers and underlying mechanisms of *Mycobacterium tuberculosis* (*M.tb*) infection, the cause of tuberculosis (TB) in susceptible individuals. TB is an important global disease, as over 2 billion people are currently infected worldwide, with an estimated 10 million new diagnoses and 1.5 million deaths in 2019 [21]. TB results from complex host-pathogen interactions, that contribute to a spectrum of TB disease forms including rapid mortality (fulminant TB), chronic disease (pulmonary TB), and more resistant forms: incipient TB, latent *M.tb* infection, and early clearance of the bacteria. Furthermore, although immunodeficiency, diabetes, and old age are known risk factors for disease progression, most patients with latent *M.tb* infection develop active pulmonary TB with no known risk factors [22]. Moreover, there exists no widely validated lung or blood biomarkers that accurately differentiate each form of TB in part because few animal models develop human-like disease phenotypes, thereby limiting translational experimental findings.

To address these limitations, we use the Diversity Outbred (DO) mouse population [23]. Each DO mouse is a heterozygous mosaic of DNA inherited from 8 founder strains [24,25]. The populations' and the individuals' genetic diversity rivals that of humans and has been used to study disease mechanisms [23–29]. When infected with *M.tb*, DO responses better emulate human TB than inbred strains [23,30–33]. We have observed a wide range of phenotypes in survival, weight change, lung granulomas, inflammatory and immune responses, which are not observed in *M.tb*-infected inbred strains [34–44]. Analogous forms of TB in humans and DO mice are shown in Table 1.

Our past work has utilized image analysis and deep learning to automatically detect histological features of *M.tb*-infected lungs of DO mice, including granulomas, cell-poor caseous necrosis, lymphocytic cuffs, macrophage-rich regions, neutrophil-rich regions, normal lung tissue, and acid-fast stained *M.tb* [35,37,45–47]. Our most recent work utilizes attention-based multiple instance learning (MIL) to classify DO mice into as "supersusceptible" (SS) and "non-supersusceptible" (nSS) with high accuracy ($91.50 \pm 4.68\%$) using only slide-level labels [48]. The work was particularly notable for its interpretability by human pathologists – examination of the MIL model "attention-weights" revealed that the model was making diagnostic decisions using a form of cellular necrosis: karyorrhectic and pyknotic nuclear debris.

We sought to develop a deep-learning method to predict gene expression of a subset of genes from WSI as an *intermediate* for classifying SS and nSS (Fig. 1). Recent studies similarly used deep learning to predict susceptibility of infectious diseases. Abdullal *et al.* compared a multilayer perceptron to traditional regression to predict susceptibility to COVID-19 [49]. Zhang *et al.* utilized genetic algorithms to develop an autoencoder to extract and cluster features related to gene expression to predict susceptibility to sepsis [50]. Shashikumar *et al.* developed an interpretable recurrent neural network to preemptively predict sepsis based on temporal features such as heart rate and arterial pressure [51]. Recent studies have also utilized deep-learning models to predict expression of disease associated genes. Levy-Jurgenson *et al.* fine-tuned a pretrained Inception-v3 to predict high and low expression of breast and lung cancer associated genes using images patches from digitized haematoxylin and eosin (H&E) slides, allowing a degree of spatial resolution for genes [52]. Dolezal *et al.* utilized a pretrained Xception to predict BRAF-RAS score from digitized H&E slides, a correlate and conglomerate of genes' expression, in non-invasive follicular thyroid neoplasms, similarly allowing for a degree of localized scoring [53]. Xu *et al.* leveraged deep learning with a novel quantitative method for measuring DNA methylation in order to predict expression of H3K4me3 [54]. Schmauch *et al.* developed a multilayer perceptron to predict RNA-seq expression of the whole transcriptome using features extracted from generic features of a pretrained ResNet-50 from digitized H&E biopsies [17].

In this present work, we developed a deep-learning method to predict gene expression of a subset of genes from WSI as an *intermediate* for classifying SS and nSS (Fig. 1). Training and validation of our model utilized a set of DO mice experimentally infected with *M.tb* from which H&E-stained lung biopsies and gene expression data (microarray) were acquired. We demonstrate that a modified version of our previous attention-based MIL model [48] that *regresses* to gene expression values is sufficiently accurate to discriminate SS from nSS

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)

better than our previous method when validated on a gradient boosted tree (GBT) classifier trained on ground-truth gene expression data.

2. Methods

2.1. Study design

Overall, our objective was to predict SS of a population of DO mice experimentally infected with *M.tb* from H&E images [30,37] using gene expression data as an intermediate. This was inspired by the high accuracy observed when predicting SS using gene expression data. However, as gene expression data is not widely available and H&E is widely available (both in our dataset and in other applications), we did not want our model to depend on it during inference – only during training. Our resulting methodology was three-fold – (1) to select a set of differentially expressed genes that accurately classifies SS (i.e., feature selection on gene expression data) using GBTs, (2) to modify our previous attention-based MIL model [51] to regress H&E images to selected genes' expression, and (3) to validate predicted gene expression data using GBTs trained on ground truth gene expression to classify SS. Specific justifications for each step are explained in respective subsections. Fig. 1 depicts a flowchart for the overall methodology. Mice and *M.tb* infection, diagnostic categories of mice, slide preparation, and digital imaging protocols are same to our previous studies [48].

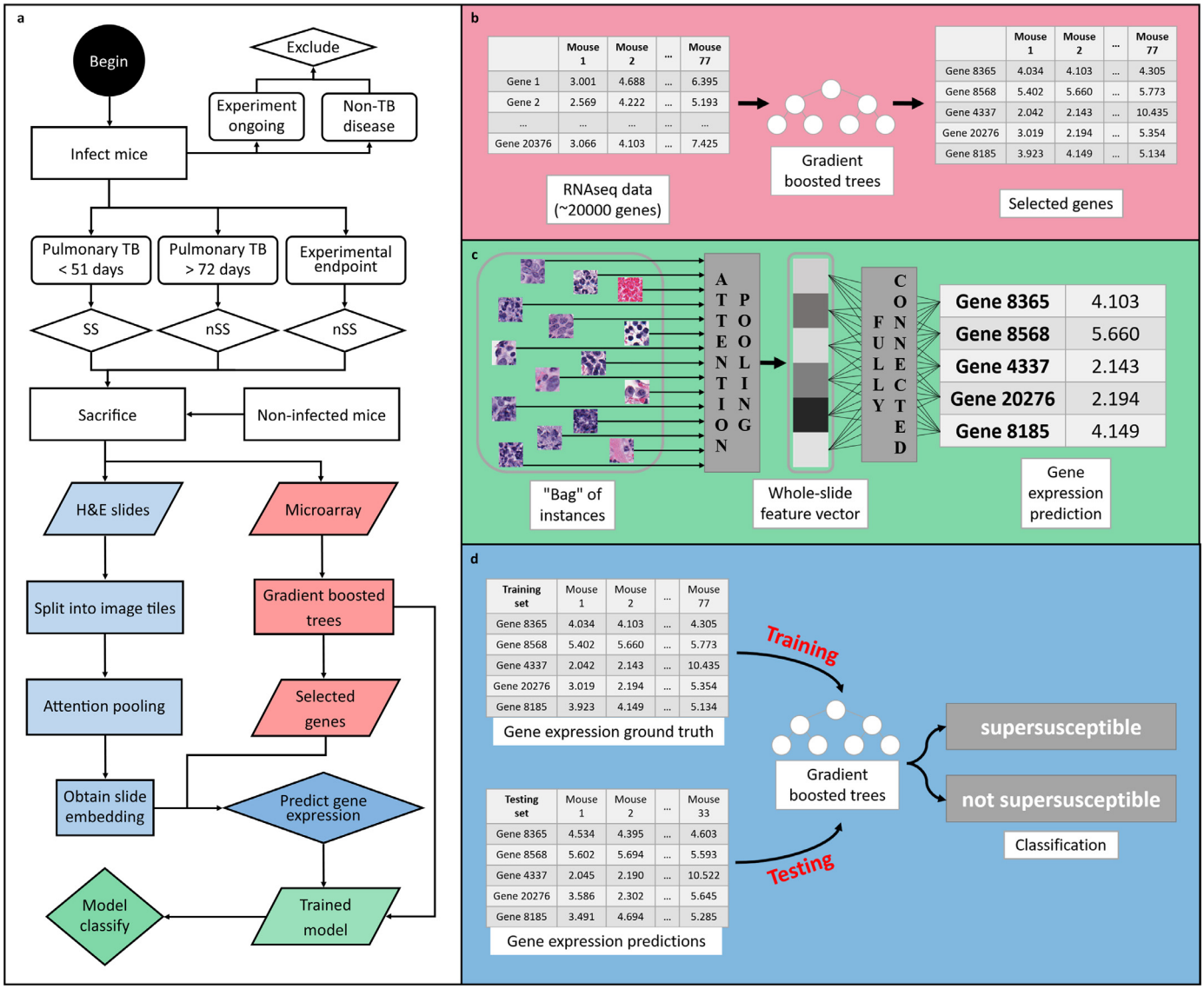


Fig. 1. Flowchart for the overall method. (a) GBTs select genes that accurately identify SS. (b) MIL predicts gene expression from H&E images. (c) Predicted gene expression validates in ground truth label trained GBTs.

2.2. Ethics statement

All ethical procedures for the study were approved by Tufts University's Institutional Animal Care and Use Committee (IACUC) protocols: G2012-53; G2015-33; G2018-33. We complied with all animal procedures.

2.3. Mice and M.tb infection

At 8-10 weeks old, female DO mice (RRID:SCR_016408) were infected with very low (~20 bacilli) dose of *M.tb* strain Erdman (ATCC strain designation 38501) using a CH Technologies nose-only system. Mice were randomized in cages prior to infection [45,48]. After infection, mice were monitored daily for health, weighed at least weekly and euthanized when IACUC-approved removal criteria were met or at the predetermined experimental time point. Euthanasia was required and performed when body condition score was < 2.0; severe lethargy was observed; or severely increased respiratory rate or effort was observed. Euthanasia by CO2 asphyxiation was the primary method, followed by vital organ removal as the secondary method. No anaesthesia was performed. Non-infected control DO mice matched for age and gender were identically housed, monitored and euthanized at experimental end points.

2.4. Diagnostic categories of DO mice

The "supersusceptible" and "not-supersusceptible" ground truth labels reflect clinical outcomes that occurs during experimental *M.tb* infection, the former characterized by morbidity and mortality within 8 weeks of infection and the latter characterized by survival longer than 8 weeks without morbidity or mortality. These phenotypes are robust and reproducible.

2.5. Slide preparation and digital images

Lungs from each mouse were inflated and fixed in 10% neutral buffered formalin, processed and embedded in paraffin, sectioned at 5μm and stained with H&E. H&E-stained glass slides were magnified 400 times and digitally scanned by Aperio ScanScope at 0.23 microns per pixel with quality factor 70.

2.6. Sample size and selection of mice for gene expression analysis

Lung samples from non-infected DO mice and *M.tb*-infected DO mice, representing a spectrum of disease, survival, and *M.tb* burden phenotypes from SS and non-SS mice were selected for gene expression profiling using microarray. The training set (Set 1) consisted of 77 samples from mouse lungs. Briefly, one lung lobe was homogenized in TRIzol and stored at -80C until total RNA extraction using Pure Link RNA mini-kits (Life Technologies, Carlsbad, CA). RNA was checked for purity and samples were analysed at the Boston University Microarray and Sequencing Resource Core Facility (Boston, MA). Mouse Gene 2.0 ST CEL files were normalized to produce gene-level expression values using the implementation of the Robust Multiarray Average (RMA) in the affy R package (version 1.36.1) and an Entrez Gene-specific probeset mapping (version 17.0.0) from the Molecular and Behavioural Neuroscience Institute (MBNI) at the University of Michigan. Array quality was assessed by computing Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) using the affyPLM R package (version 1.34.0) and by normalizing the CEL files using Expression Console (build 1.4.1.46) and the default Affymetrix probesets to compute the Area Under the [Receiver Operating Characteristics] Curve (AUC) metric. All microarray processing was performed using the R environment for statistical computing (version 2.15.1). Principal component analysis was run on the resulting data matrix (Supplementary Fig. 1).

Near the end of our model development, the testing set (Set 2) – consisting of 33 lung samples from DO mice with susceptibility labels, lung gene expression data, and digitized H&E slide images of lungs – became available. The slides and gene expression data from Set 2 were imaged in different batches, and RNA was extracted, processed, and analysed in a separate batch from Set 1. Thus, Set 1 (77 samples) and Set 2 (33 samples) were independent and the latter used as an external testing set for model development.

3. Model description

3.1. GBTs for feature selection

Set 1 was relatively imbalanced in terms of SS (n=23) and nSS mice (n=54). Furthermore, experimental conditions varied across subsets of mice within Set 1 (i.e., different initial *M.tb* doses and batch effects of microarrays). Finally, our experimentally infected DO mice have over 20,000 genes for which expression data was available and the majority of these expression patterns are not predictive of the SS phenotype. To reduce the number of genes and identify genes predictive of SS while simultaneously accounting for the aforementioned deficits in Set 1, we posed GBTs as a feature selector, in which genes were features, gene expression values were feature values, and SS was the outcome being predicted. Specifically, we utilized the Xgboost implementation [55] as a novel method for feature selection [56]. Set 2 was similarly imbalanced (11 SS; 22 nSS), contained mice across experiments (i.e., different initial *M.tb* doses), and consisted of gene expression values derived from a distinct microarray batch.

GBTs are an ensemble of decision trees that operate on the principle that adding additional trees to the ensemble should emphasize data points that are incorrectly classified prior to adding additional trees. For example, if a single tree misclassifies some subset of training data, then the next tree added to the ensemble should focus on the misclassified training data. Boosted trees accomplish this by giving greater weight to misclassified training samples while computing overall error. However, GBTs instead construct new trees by directly using the error (called a residual) from the current state of the ensemble. Thus, the error of the current state of the ensemble is minimized rather than the overall error after adding an additional tree.

Hyperparameter selection was carried out for 100 iterations of a 10-fold cross-validation on Set 1 (details in Supplemental Methods). For each trial, the unique set of genes utilized in the fitting of the model were recorded. Following the gene selection process, the top 1 to top 15 most frequent genes (across all 100 trials) were selected, and an Xgboost model was fit to predict SS for each set of genes using leave-one-out cross-validation to determine how many genes to use and to ensure model performance did not degrade. This leave-one-out procedure was carried out 15 times due to the stochastic nature of hyperparameter selection. Fig. 2 depicts the process of feature (gene) selection.

3.2. Attention-based MIL for WSI Processing

Three primary challenges arise when processing WSIs using deep learning. First, deep learning models require strong labels [57]. This usually takes the form of hand-drawn annotations on images. Manual annotations are time-consuming to create and sometimes cannot be known. Strong labels thus are contrasted with weak labels, which are assigned to a whole image rather than its individual components. A second primary challenge is that conventional deep learning models are incapable of quickly processing large images. Normally, this problem is resolved by resizing images but cannot apply for WSI processing, as fine (often pertinent) details such as cells, location information, and tissue-level microanatomy are lost. As a solution, deep learning models tend to sample smaller image crops ("tiles") from the WSI. But this is computationally expensive, as WSIs may

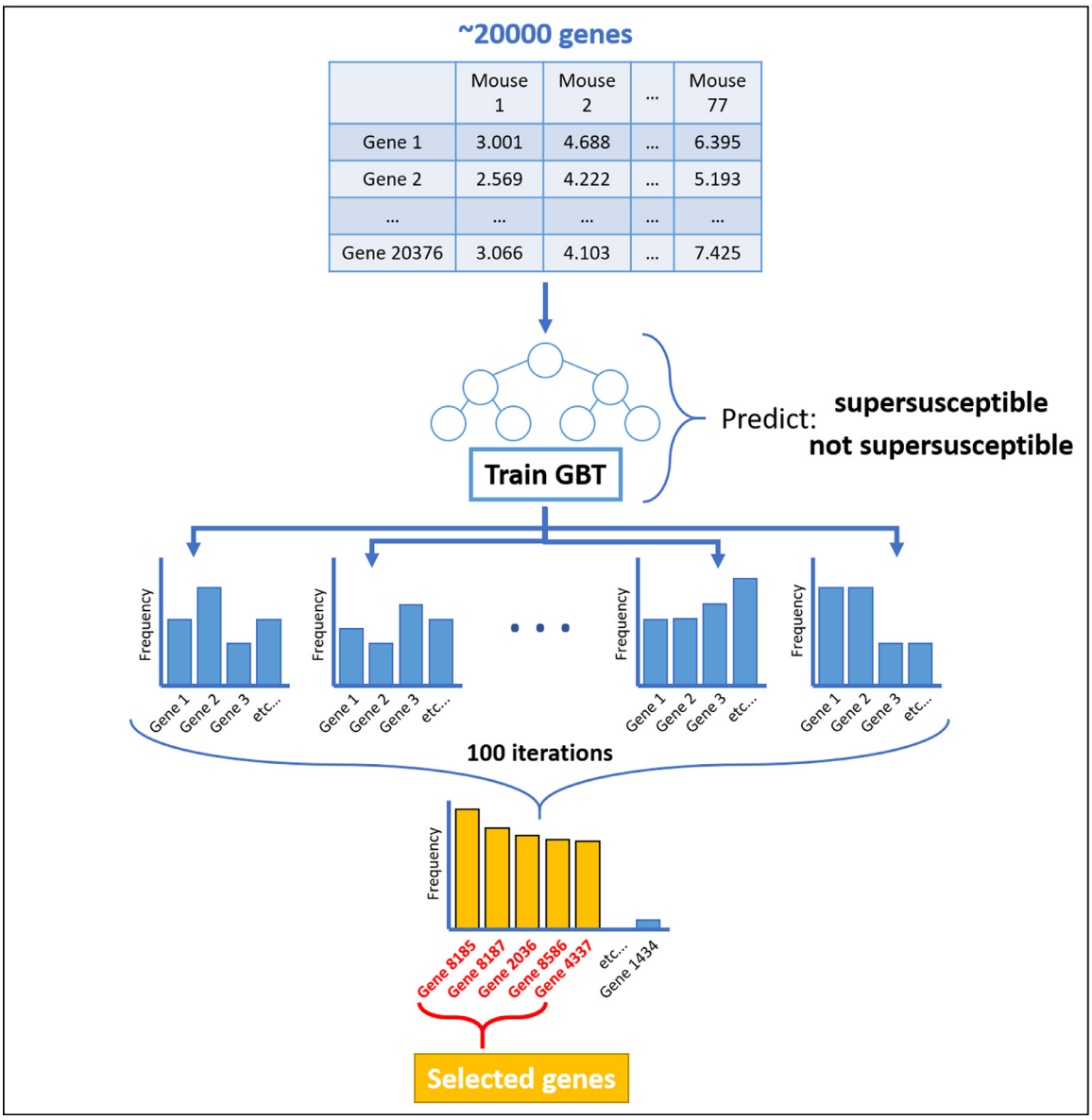


Fig. 2. Selection of genes. ~20000 genes are used as features to predict SS using Xgboost. This is iterated 100 times, while the genes utilized by the model are recorded. Then, the five most frequently occurring genes were utilized for subsequent experiments.

contain hundreds of thousands of tiles. Finally, a third primary challenge is that deep learning models are limited by their interpretability [1]. This 'black-box' nature of deep learning limits its applicability in research and medicine, as both scientists and clinicians must know how decisions are made before informing a biological mechanism or clinical decision [1,47]. Attention-based deep MIL [48,58] offers a solution to each of these problems in deep learning when processing WSIs. A background on MIL and details regarding its attention-based implementation can be found in the Supplementary Methods. When the MIL paradigm is applied to WSI processing, bags are analogous to WSIs and instances are analogous to tiles taken from slides [4].

In our specific case, labels are analogous to SS or gene expression data. This is because the areas of tissue indicative (i.e., predictive) of SS are certainly implicit but not explicitly known. Similarly, gene expression of localized lung regions is not explicitly known but is implicit by the known overall expression of the whole tissue. When posed as such, MIL resolves the first two problems described previously. First, weak labels are generally easily available for WSIs, such as a disease state or diagnosis. Specifically, both SS and gene expression are weak (and not strong), as they apply to the whole tissue. Second, tiles taken from slides (when subsampled) make processing WSIs quicker without the need for resizing (although there are issues with subsampling – see next section). The third problem, interpretability, is resolved via attention pooling (Supplemental Methods). Briefly, attention pooling automatically learns how to weight tiles according to their pertinence to the overall target, in our case gene expression. Thus, the magnitude of these weights gives insight into what the model is paying attention to, yielding interpretability.

3.3. Sampling of WSIs

Ideally, slides would not need to be sampled (i.e., tiles taken as an instance of the slide), and every part of the tissue would be included in the training of the model. However, due to the sheer size of WSIs (~100,000 × 100,000), this would result in hundreds of thousands to millions of instance tiles depending on their size. For a relatively shallow feature extractor and attention mechanism (as in Supplementary Table 3), this number of instances far exceeds the limitations of modern GPUs (with ~16GB memory), as gradients for each instance need to be recorded during training. This limitation is akin to batch size

limitation in training conventional convolutional neural networks. In previous experiments [48] with instance sizes of 32 × 32 pixels and 256 × 256 pixels, the resulting limitation was around 5000 and 100 instances, respectively. These experiments also demonstrated that the spread of instances across the slide was more important than tile size. Thus, experiments in this study utilized 5000 32 × 32 pixel tiles to maximize spread across the slide. Yet, simply sampling 5000 tiles from each slide proved fruitless (see Results), so the sampling procedure was altered. Instead, for each slide, 100,000 tiles were extracted at random. Then during training of the MIL model, a random subset of instances was selected in order to not exceed GPU memory limitations. We thought this to be an appropriate alternative because over several iterations, each instance should *eventually* be used. As a comparison, 75 bags were created for each slide by randomly sampling 2500 32 × 32 pixel tiles for each bag. This method effectively functioned as a manner of data augmentation, taking the set of 77 mice to 5775 mice. These sampling methods are depicted in Fig. 3.

3.4. Attention-based MIL to predict gene expression data from WSIs

After the selection of candidate genes and bags from slides, models were trained using an attention-based MIL model to predict gene expression values. Our implementation [48] mapped bag-level feature vectors to a single value using a fully connected layer followed by sigmoid activation (which constrains the output to between 0 and 1 – i.e. a probability). A threshold was then applied to perform classification, as in Eq. (1).

$$\begin{cases} 1 \ if Prob \geq 0.5 \\ 0 \ if Prob < 0.5 \end{cases} \quad (1)$$

$$\widehat{Y} = zW \quad (2)$$

To perform regression, the sigmoid activation and thresholding steps were removed and replaced with a fully-connected layer. Thus, the regression step takes as input the bag-level feature vector (z) multiplies it by a matrix of learned weights (i.e. a fully connected layer) and outputs predictions for gene expression values, as in Eq. (2). In addition, the initial feature extraction layers and attention module remained the same. As we eventually utilized only five genes, the ultimate fully-connected layer, W, was a 90 × 5 matrix, and $\widehat{Y}$ was a 5 × 1 vector of predicted gene expression values. The resulting model is summarized in Supplementary Table 1. Each model was trained for 200 epochs using the Adam optimizer with a learning rate of 0.0003, weight decay of 0.0005, betas of 0.9 and 0.999, and training error cut-off of 0.1. Mean-squared error was used as a loss function for the regression problem. Training was halted if the validation set loss did not improve for more than 15 epochs. The model was saved each time the validation loss decreased. Models were only fit to mice in Set 1. Fig. 4 depicts WSI gene expression prediction.

3.5. Assessment of gene expression prediction

With models trained to predict the top five gene expression values, an assessment of how accurate these predictions was needed. During experimentation, we often examined mean absolute error over the distribution of gene expression values for single genes in order to see which range of ground truth gene expression values resulted in the most error. Although this was effective in comparing model to model, it could not be used as an absolute measure for accuracy because it is not known how much error is acceptable in predicting gene values. An error of 0 is the goal, but an error of 1 is perhaps too much given that the range of gene expression values of the gene we examined ranged from 2 to 12. Furthermore, gene expression values are log-transformed, so larger errors are more acceptable for larger values. As a solution and to compare the MIL method to our previous method [48], we decided to pass predicted gene expression values through Xgboost model trained on ground truth gene expression data to predict SS. In this manner, the accuracy of predicting SS was used as a proxy for the accuracy of gene expression values

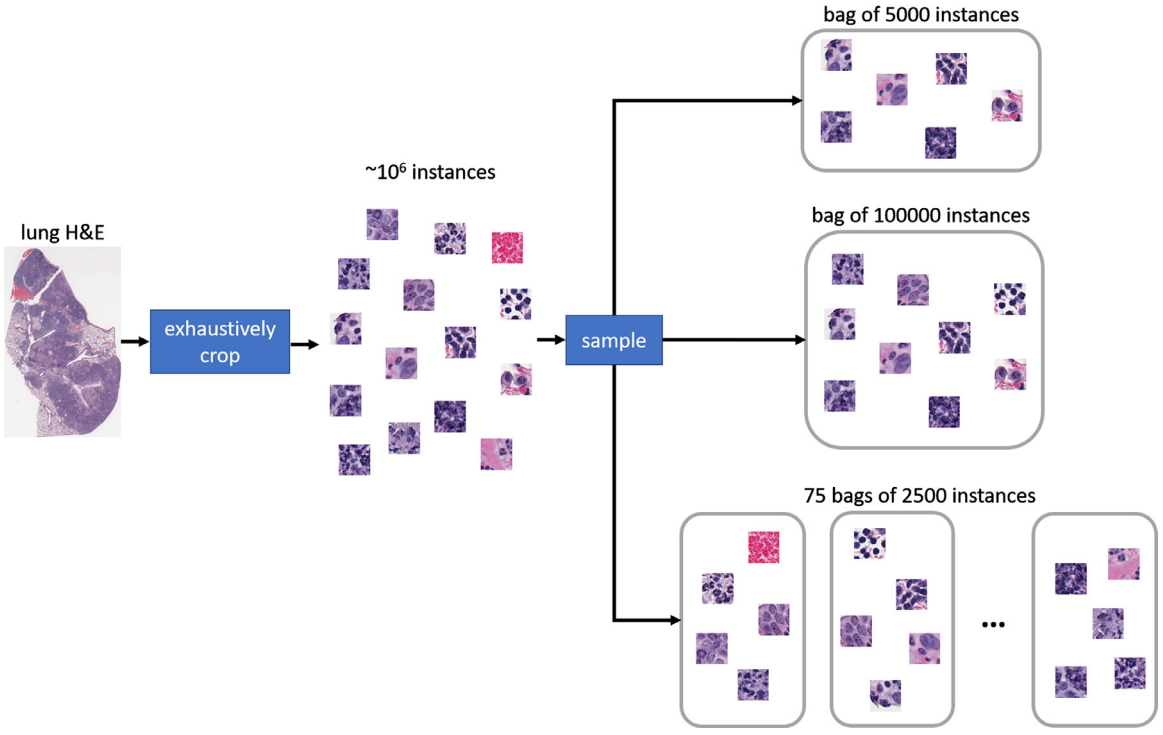


Fig. 3. Sampling methods. In the first sampling, method 5000 random tiles are selected for a slide. In the second, 100,000 random tiles are selected for a slide. In the third, 2500 random tiles are selected for a slide 75 times.

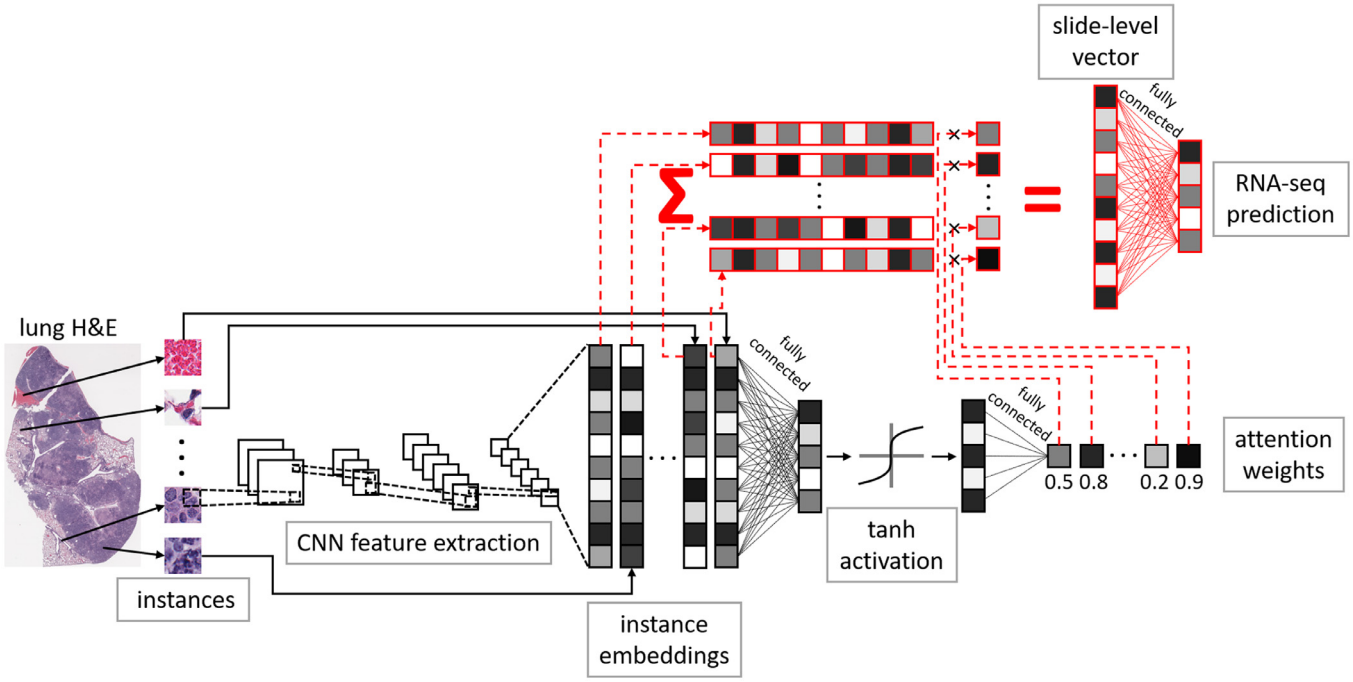


Fig. 4. Gene expression prediction via an attention-based MIL model. Instances (32 × 32 pixels) are randomly extracted from the lung H&E sections of each mouse and passed through a feature extractor to yield instance embeddings. Embedding are passed through the attention module to yield an attention weight and thusly scaled and summed to a slide-level vector. Gene expression is predicted from this slide-level vector using a fully-connected layer.

predicted by the attention-based MIL model. Xgboost models were trained using the same hyperparameter search as previously described in the Methods section. Their specificity and sensitivity were compared to our previous model, which utilized attention-based MIL to directly predict SS from slides.

3.6. Model summary

Set 1 was used to select genes from 20000 using Xgboost as a feature selector. Next, the MIL model was trained using a 12-fold Monte-Carlo cross-validation on Set 1 mice using various bag sampling strategies to predict expression values of all five genes. These same folds were utilized to cross-validate the Xgboost model to predict SS using the selected genes. Predicted gene expression values from the MIL model were passed through their respective Xgboost models to assess the accuracy of gene expression prediction and to compare to slide sampling methods. Finally, the independent Set 2 was passed through each fold of MIL model to predict the expression of the five selected genes. These predictions were passed through an Xgboost model (trained on Set 1) to predict SS as a proxy for assessing gene expression prediction accuracy and to compare models.

4. Analysis

4.1. Statistical methods

GBT performance was evaluated using overall accuracy, sensitivity, and specificity of a ten-fold cross-validation for the full set of genes and subsequent leave-one-out cross-validation using selected genes as described above (R 3.6.3). Accuracy of gene expression prediction was evaluated using correlation and cosine similarity for each sampling method using a 12-fold cross-validation as described above (MATLAB 2020a). Finally, the performance of the H&E to gene expression model was evaluated using accuracy, sensitivity, and specificity using the same 12 folds on a GBT model trained on ground truth gene expression values (R 3.6.3). This procedure was carried out for Set 1 and then validated using Set 2.

4.2. Role of funding source

The funders had no role in study design, data collection, data analysis, interpretation, or writing of the report.

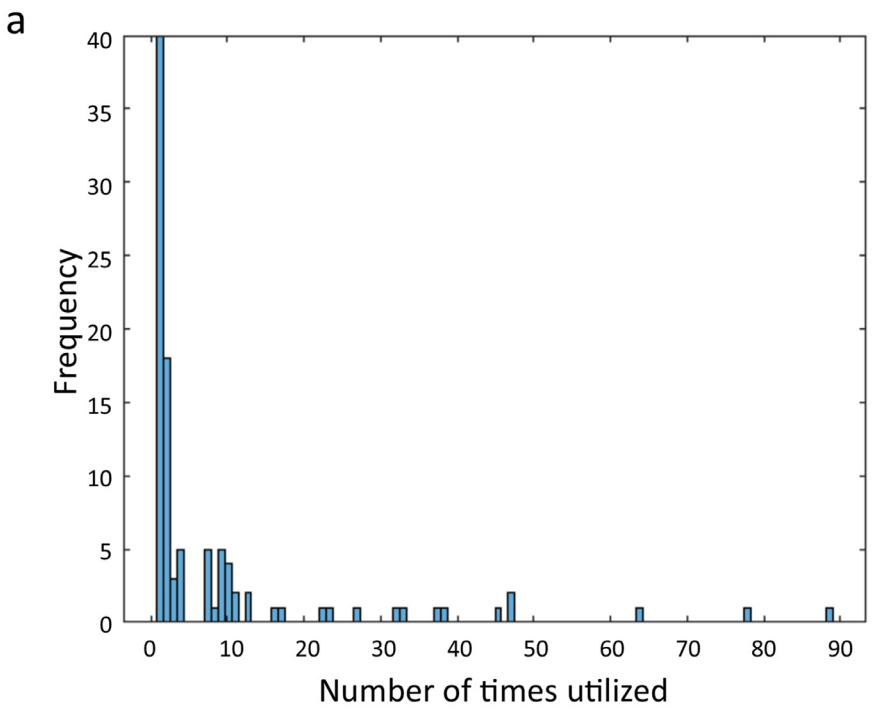
5. Results

5.1. GBTs for feature selection

A total of 896 genes were utilized by the GBT classifiers, 100 of which were unique. 40 of these genes were utilized in one trial, 60 in more than one trial, and one gene (*serpina3n*) in 89 trials (Fig. 5a). The testing sensitivities and specificities (mean±std) for classifying SS and nSS in these trials were 97.00±8.21 and 92.00±15.67. Leave-one-out cross-validation of the top genes (Fig. 5b) resulted in high sensitivities and specificities (Fig. 5c). Given that the sum of mean sensitivity and specificity for the top 5 genes was the highest, the remainder of experiments utilized the top 5 genes – *serpina3n, ifitm6, serpina3m, cxcr2*, and *ms4a8a*. The leave-one-out cross-validation sensitivity and specificity utilizing these top 5 genes was 97.37±1.57 and 98.06±2.57. The variation derives from the hyperparameter selection.

5.2. Attention-based MIL to predict gene expression data from WSIs

Following sampling of each slide, an attention-based MIL model was trained to predict gene expression values for the top 5 genes via 12-fold Monte-Carlo cross-validation. The Pearson correlations between predicted and ground truth gene expression values across Set 1 folds are reported in Table 2 for each sampling method – 5000 instances per slide [48], 100,000 instances per slide [48], and the proposed 75 bags of 2500 instances per slide – reported in respective triplets. To come to a consensus gene expression prediction for the 75 bags of 2500 instances, an average was taken across all 75 output vectors. Overall, correlation is high for this sampling method and exceeds the correlation yielded by simply sampling 5000 instances per slide [48]. In these contexts and throughout the Results, ground



b

Gene symbol	Frequency
serpina3n	89
ifitm6	78
serpina3m	64
cxcr2	47
ms4a8a	47
orm3	45
mmp8	38
retnlg	37
ly6g	33
mt1	32
mt2	27
gm14548	23
Mmp9	22
il1r2	17
f13a1	16

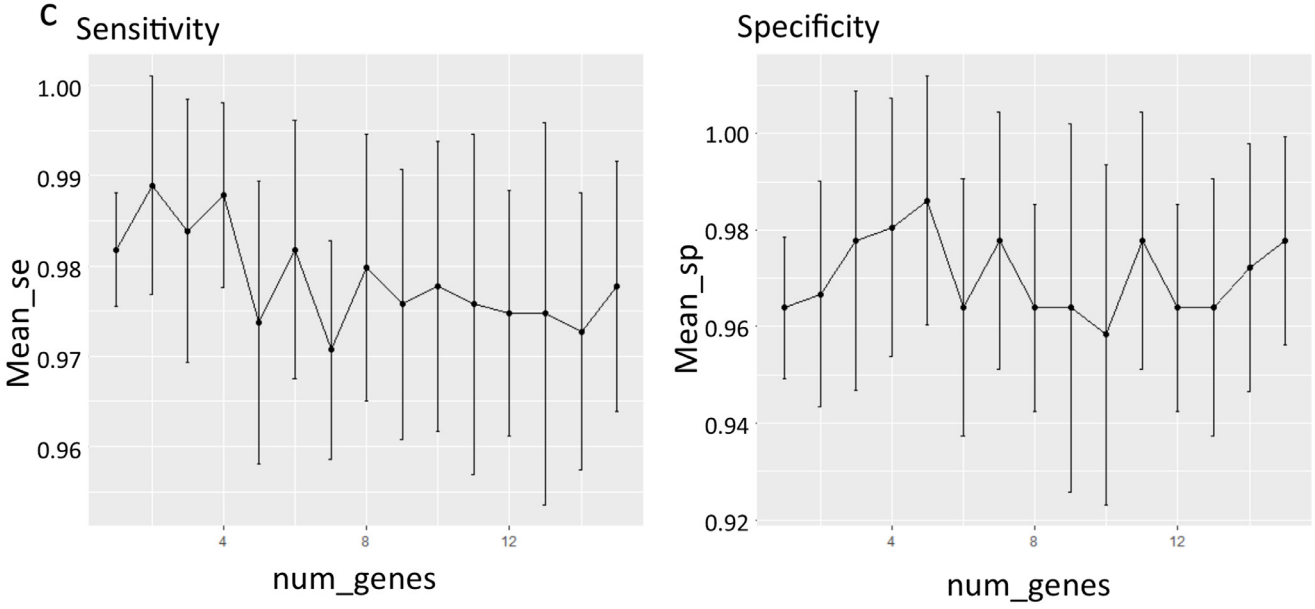


Fig. 5. Results of GBTs for feature selection. (a) Distribution of gene utilizations – only one gene was utilized 89 times. (b) Frequency of each gene use (top 15 shown). (c) Results of cross-validation on top-n genes (x axis) and their sensitivities and specificities (y-axis).

truth gene expression refers to the raw log transformed values reported by microarray analysis, and predicted gene expression refers to the attention-based MIL model output based on H&E. Training and validation refer to mice in Set 1 (n = 77) utilized for model development. Testing refers to mice in Set 2 (n=33), an external test meant to assess out-of-sample performance of each model.

Though correlation is high, it alone does not indicate that the values are the same. Thus, we report the cosine similarity of predicted

Table 2
Ground truth and predicted gene expression correlation.

	Training(Set 1)	Validation(Set 1)	Testing(Set 2)
serpina3n	0.06 | 0.26 | **0.86**	0.18 | 0.59 | **0.69**	0.16 | 0.13 | **0.82**
ifitm6	0.12 | 0.29 | **0.90**	0.25 | 0.60 | **0.84**	0.15 | 0.16 | **0.82**
serpina3m	0.07 | 0.25 | **0.85**	0.13 | 0.36 | **0.63**	0.13 | 0.14 | **0.81**
cxcr2	0.09 | 0.14 | **0.81**	0.29 | 0.50 | **0.69**	0.25 | 0.24 | **0.65**
ms4a8a	0.12 | 0.24 | **0.89**	0.34 | 0.73 | **0.81**	0.22 | 0.22 | **0.84**

Table 3
Ground truth and predicted gene expression cosine similarity.

	Validation(Set 1)	Testing(Set 2)
5000	0.9961	0.9957
100000	0.9963	0.9958
75 bags of 2500	**0.9980**	**0.9970**

Table 4
Performance predicting high/low gene expression.

	Training(Set 1)	Validation(Set 1)	Testing(Set 2)
serpina3n	0.9220/0.1103 0.8571/0.2273 **0.9701/0.7958**	0.8983/0.1200 1.0000/0.0000 **0.9661/0.8000**	0.8826/0.1212 0.7682/0.2545 **0.8977/0.8030**
ifitm6	0.9120/0.1765 0.8750/0.3043 **0.9552/0.8896**	0.8571/0.1786 1.0000/0.5000 **0.9286/0.8214**	0.8587/0.1500 0.7565/0.2300 **0.8188/0.8917**
serpina3m	0.7514/0.2737 0.7234/0.5833 **0.8951/0.8248**	0.7193/0.4074 0.5000/1.0000 **0.8596/0.8519**	0.7542/0.3397 0.5750/0.5308 **0.8792/0.8846**
cxcr2	0.9646/0.0764 0.9184/0.0455 **0.9679/0.6867**	1.0000/0.0800 1.0000/0.0000 **0.9492/0.7600**	0.9400/0.1250 0.9480/0.1625 **0.8067/0.4479**
ms4a8a	0.9809/0.0417 0.9388/0.0455 **0.9621/0.7766**	0.9833/0.0000 1.0000/0.0000 **0.9667/0.8333**	0.9625/0.0321 0.9200/0.0692 **0.9917/0.7051**

and ground truth gene expression values by taking both as a vector of values in Table 3 – 5000 instances per slide [48], 100,000 instances per slide [48], and the proposed 75 bags of 2500 instances per slide. Overall, the sampling method using 75 bags is more accurate than dynamically sampling from 100000 instances or the 5000 instances of our original method. Cosine similarity is defined as the inner product of two vectors divided by product of their magnitudes. In essence, it reflects the angle between two vectors – the more they are aligned, the higher the value (between 0 and 1).

Finally, we report model performance in terms of predicting high/low gene expression for each gene in Table 4. As in Table 2, we report results per gene in pairs for Set 1 and Set 2. Each cell contains three rows, where each corresponds to a sampling method (5000 instances per slide [48], 100,000 instances per slide [48], and the proposed 75 bags of 2500 instances per slide). Each member of a pair refers to high/low classification performance. Overall, the proposed sampling method outperforms conventional sampling methods.

5.3. Assessment of gene expression prediction

Predicted gene expression values for the validation set of each fold were then passed through respective Xgboost models that were previously trained on ground truth gene expression values of the training set of each fold. The overall cross-validation sensitivity and specificity for predicting SS are reported in Table 5 for each sampling method – 5000 instances per slide [48], 100,000 instances per slide [48], and the proposed 75 bags of 2500 instances per slide – expressed in sensitivity/specificity pairs. Clearly, the sampling method using 75 bags of 2500 instances per slide is superior in terms of accuracy.

6. Discussion

Our results indicate that our attention-based MIL model can be extended to accurately regress a subset of gene expression values. This is important, as recent methods have framed MIL solely as a classification problem in processing WSIs [4,58–62]. This makes sense in the context of computational pathology, as most popularly, the disease of interest is cancer, and the desired outcome is grade (or multi-class classification), or the existence or non-existence of a certain disease is desired (binary classification). However, there are continuous clinical intermediates, gene expression being just one of them. These intermediates are essential, as they provide the information clinicians need to make informed clinical decisions. The proposed MIL model not only opens an avenue for predicting other gene expression values (i.e., other gene products for other diseases) but also other continuous intermediates in medicine, such as tumour microsatellite instability and immune infiltration. Moreover, we have shown that more than one continuous variable can be accurately predicted (i.e., multi-regression). Thus, our model is not limited to predicting only one value.

Table 5
Sensitivity/specificity of SS classification using predicted gene expression values.

	Validation(Set 1)	Testing(Set 2)
5000	00.00/100.00	00.00/100.00
100000	21.74/90.74	00.00/95.24
75 bags of 2500	**87.50/95.00**	**88.03/92.52**

Furthermore, we have shown that whole-slide MIL datasets may be augmented not in traditional sense (by which images are colour jittered, rotated, flipped, blurred, and warped) but in a "bag" sense in which multiple bags may be drawn from the same slide. This latter point is supported by the higher correlation (Table 2), cosine similarity (Table 3), performance (Table 4), and resulting accuracy (Table 5) in identifying SS mice. This is not only because it is a viable data augmentation step but mostly because WSIs are huge. As a result, deep learning methods rarely process them as whole. Consequently, most methods for processing WSIs focus on processing small regions, namely high-power fields (HPFs). Though these solutions circumvent the problem of being able to process huge WSIs, they do not fully address the problem of how to sample WSIs. This is a well-known problem in the pathology community. Pathologists often sample regions of the slide when making a diagnosis, and these regions often vary by pathologist, thus the analyses and decisions derived from them [1,63]. MIL methods have been proposed as promising solutions to the original problem – by allowing the processing of a whole-slide for clinical tasks – and have seen many successes [4,61]. Yet, there are still hardware limitations, as current hardware is only capable of fitting a limited number of instances per slide into memory. The proposed sampling methodology overcomes this barrier by decomposing a slide into multiple bags, thereby processing the whole slide as a result. The slide is well-represented with hardware-permitting magnitudes of instances.

Given that the best performing sampling method (75 bags per slide) dramatically increased the size of the dataset, the overall training time similarly increased. Though this sampling method yielded more accurate results, a larger dataset (i.e. with more slides) would similarly require longer training times. However, if the proposed sampling method is viewed as a data augmentation step (as discussed previously), it is possible that utilizing 75 bags per slide would not be necessary for a dataset with more slides. We will explore this possibility on larger datasets in the future.

Recently, investigators used a different a deep learning approach to predict gene expression data from WSIs using a multi-step process [17]. Briefly, their model was trained on WSIs of tumour biopsies, with many origins, to predict the full gene expression data, not just specific genes. WSI tissue area was spatially clustered using k-means (k=100), and 8000 tiles (224 × 224 pixels) were extracted from each slide. These tiles were passed through Resnet50 [64], and their features extracted into a 2048-dimensional vector. These vectors were then averaged into 100 vectors based on the initial spatial clustering, dubbed "super-tiles". A multi-layer perceptron was then trained on these supertiles to predict gene expression data. During training, only the top k [1,2,5,10,20,50]· 100 tiles in terms of predicted gene expression are selected and a weighted mean computed of their predictions, giving more weight to higher valued predictions. Finally, the

model is fine-tuned for specific organs using the 8000 normal tiles, again sampling the top k [10,20,50], 100, 200, 500, 1000, 2000, 5000 during training. There are three limitations to this method and its results.

First is their initial clustering step. This step attempts to cluster homogenous areas of the slide such that their resulting feature vectors are relatively similar. Thus, when averaged, the resulting "supertile" remains on the same manifold as the original tiles. The authors probably did this to reduce the effective number of samples to decrease training time. Clustering based on slide coordinates intuitively makes sense, as local regions tend to be more similar than distal regions. However, this is not always the case. As tissue structures on WSIs may be relatively complex in terms of shape, clustering based on coordinates will result in clusters that do not necessarily encompass homogenous tissue areas. Though some tissue structures will fit this shape, many will not, such as layers of a colon or bladder. Thus, homogenous areas of tissue are not guaranteed to be in the same spatial cluster. Thus, when extracted features from tiles selected from this heterogeneous cluster are averaged, the resulting feature vector may not lie on the original manifold of the set of all tiles [65]. For example, in three dimensions, extracted features from tiles of one tissue type may lie on a plane, while another tissue type may lie on a different plane. If all these points are averaged, unless the planes are the same plane, the resulting feature vector will not be on either plane. Both spatial positions, as well as value (i.e., the value of the features coming from local regions), are important when considering clustering for such a purpose. This is evidenced by such methods as mean-shift segmentation, which considers both spatial coordinates as well as feature values [66]. Our proposed sampling method, based on visual evidence, guarantees that homogenous regions of the slide receive the same attention weight; thus, a true heterogeneous sampling can be achieved.

Second is the choice of the number of clusters as well the k's (number of top tiles or supertiles to consider when computing loss) during the training phase. The authors did not mention what criteria led to the selection of any of these parameters. It is probable that these parameters need to be tuned, which in the context of deep learning may take exhaustive computational resources. In the case of the proposed MIL model, these parameters need not be chosen, as the attention weighting mechanism automatically learns the weights necessary for combining embedded instances.

Third, the result of our correlation seems to be superior to those presented in [17]. Their reported correlations ranged from 0.11 to 0.47, which is weak. This indicates that the gene expression predictions were relatively inaccurate compared to the ground truth gene expression. Our experiments on our best sampling method yielded 0.63 to 0.90, indicating relatively moderate to high correlation. Though the results presented here only predicted five gene expression values compared to their 28,334. However, in preliminary experiments, the proposed MIL model was mapped to 1000 random genes rather than just the top 5. The resulting cosine similarities were on average 0.9959 and Pearson correlations on average 0.59. This suggests that the proposed MIL model may generalize to any number of genes.

The proposed methodology has limitations. Our work is based on a relatively small dataset that could limit its applicability to other disease models. However, the MIL method does generalize to an external dataset of microarray data, which was produced from a separate experimental infection, different generation of DO mice, and separate microarray batch. Finally, when applied to a larger number of genes, this generalization still holds. Therefore, we believe that the relatively small size of the dataset is not a significant problem.

The translational relevance of the current work derives from the insight gained from intensive study the lungs, the primary site of TB. Lung tissues are not readily available from human TB patients before or after death, and the treatments patients receive alter the tissue and cellular responses. Therefore, to obtain mechanistic insight into TB granuloma structure and function, we must model infection using an *in vivo* animal model that genotypically and phenotypically resembles humans. The DO mouse population is the only mouse population that captures human genetic diversity and can be used to identify genes and polymorphisms that contribute mechanistically, which is a goal of our research. No other Outbred animal model (e.g., non-human primates, rabbits, guinea pigs) can be used to address these knowledge gaps. The transcripts highly expressed in lungs of *M.tb*-infected SS DO require further study. Roles for 4 of the top 5 transcripts (*serpina3n, ifitm6, serpina3m*, and *ms4a8a)* are unknown in TB. *serpina3* transcripts encode for protein molecules in cell activation, signalling, and metabolism, and are elevated in brain tissues of patients with TB meningitis [67] but mechanisms and consequences are not known for TB. *ifitm6* transcripts encode for an interferon-inducible transmembrane protein of innate immune cells such as macrophages [68] and may generally regulate generation of T cell-mediated immunity [69] but this is unproven speculation. *ms4a8a* transcripts encode for membrane-spanning plasma proteins reported highly expressed in alternatively activated macrophages and dendritic cells in homeostasis [70] and autoimmunity [71]; and again is unknown in TB. Of the top 5, *cxcr2* has been investigated in context of TB with 18 papers retrieved from a PubMed search on 3/24/2021. *cxcr2* encodes for a chemokine receptor which recruits neutrophils and macrophages via ligands CXCL1, CXCL2, CXCL5. *cxcr2* expression is high due to many of these cells in the lungs, or high levels expressed by few individual cells (or both). This is consistent with the TB granuloma phenotypes we previously described [37]. In other models of inflammation, cell-specific deletion of *cxcr2* in neutrophils protects hosts against brain-damaging inflammation [72]. Roles (known and hypothesized for CXCR2 and its ligands CXCL1, CXCL2, and CXCL5) in TB in humans and experimental animal models have recently been reviewed [73].

Here, we presented a deep-learning model to predict SS of DO mice to *M.tb* infection using gene expression data as an intermediate modality from histology images. Overall, on our selected gene set, it is accurate, showing relatively high positive correlations with ground truth gene expression values on both cross-validation and external testing sets. Furthermore, when predicted gene expression values from validation and testing sets are passed through an Xgboost model trained on ground truth gene expression training values, the resulting classification accuracy in identifying SS mice is high. Finally, the resulting accuracy is higher than that of our previous method. In future studies, we will explore a promising biological application of the model in which expression of single genes may be localized to specific anatomical locations – "virtual staining" – using attention weights in a similar manner to our previous work. This could generate novel hypotheses regarding molecular mechanisms and could be used to infer functional differences in lesions with similar morphologies. We will additionally examine prediction of expression of a larger number of genes.

The overall methodology, from gene selection to target a specific outcome, to predicting gene expression values, to again predicting an outcome based on predicted gene expression values, we believe serves as a computational template for gene expression panels. As performing a full transcriptomic profile is not practical in clinical settings, only disease-specific genes are profiled. This aspect of gene expression panels is modelled by the proposed GBT feature selection method (bound to a specific disease) and by the proposed MIL regression method to gene expression. These profiles in the clinic then inform disease-specific clinical decisions, in our case supersusceptibility to tuberculosis. When trained for specific clinical outcomes, the overall methodology depends on H&E and is economical, efficient and deterministic in comparison to gene expression panels.

Declaration of Competing Interest

Mr. Tavolara has nothing to disclose. Dr. Niazi has nothing to disclose. Ms. Ginese has nothing to disclose. Dr. Gower has nothing to disclose. Dr. Beamer has nothing to disclose. Dr. Gurcan has nothing to disclose.

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 and verifying the underlying data. ACG contributed to review and editing of the manuscript and data analysis. GB contributed to data collection, data curation, formal analysis, conceptualization, funding acquisition, project administration, verifying the underlying data, 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. All authors read and approved the final manuscript.

Acknowledgements

We thank Julie Tzipori, Curtis Rich, Donald Girouard, and Sam Telford III for services in the New England Regional Biosafety Laboratory at Tufts University Cummings School of Veterinary Medicine, North Grafton, MA. Frances Brown, Linda Wrijil, and Sarah Ducat provided histology services at Tufts University's Cummings School of Veterinary Medicine. All microarray protocols were carried out by the Boston University Microarray and Sequencing Resource (BUMSR) core facility, and we thank Eduard Drizik of the BUMSR for the initial analysis of microarray data. We also acknowledge 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, National Cancer Institute, Bethesda, MD is acknowledged for digital slide scanning using Aperio ScanScope. Funding support was provided by NIH R21 AI115038; NIH R01 HL145411; NIH UL1-TR001430; and the American Lung Association Biomedical Research Grant RG-349504.

Data sharing statement

All data and code utilized for development and validation is available at github.com/cialab/image2gene 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.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103388.

References

[1] Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019;20(5):e253–e61.
[2] Sornapudi S, Stanley RJ, Stoecker WV, Almubarak H, Long R, Antani S, et al. Deep learning nuclei detection in digitized histology images by superpixels. J Pathol Inform 2018;9:5.
[3] Li C, Wang X, Liu W, Latecki LJ. DeepMitosis: mitosis detection via deep detection, verification and segmentation networks. Med Image Anal 2018;45:121–33.
[4] 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.
[5] Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype and intrinsic subtype. NPJ Breast Cancer 2018;4(1):30.
[6] Papanastasopoulos Z, Samala R, Chan HP, Hadjiiski L, Paramagul C, Helvie M, et al. Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. SPIE Med Imaging 2020;11314:113140Z.
[7] Anand D, Kurian NC, Dhage S, Kumar N, Rane S, Gann PH, et al. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. J Pathol Inform 2020;11(1):19.
[8] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24(10):1559–67.
[9] Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis Oncol 2020;4(1):14.
[10] Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. Gastroenterology 2020;159(4):1406–16 e11.
[11] Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25(7):1054–6.
[12] Jain MS, Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat Mach Intell 2020;2(6):356–62.
[13] Zhang H, Ren F, Wang Z, Rao X, Li L, Hao J, et al. Predicting tumor mutational burden from liver cancer pathological images using convolutional neural network. In: Proceedings of the IEEE international conference on bioinformatics and biomedicine (BIBM); 2019. p. 920–5.
[14] Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. Nat Rev Genet 2013;14(1):23–34.
[15] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21(5):631–43.
[16] Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome profiling in human diseases: new advances and perspectives. Int J Mol Sci 2017;18(8):1652.
[17] Schmauch B, Romagnoni A, Pronier E, Saillard C, Maille P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun 2020;11(1):3877.
[18] Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet 2016;17(5):257–71.
[19] Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, et al. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. Int J Mol Sci 2017;18(2):308.
[20] Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet 2019;20(6):341–55.
[21] (WHO) WHO. Tuberculosis 2019 [05/20/2020]. Available from: https://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.
[22] (WHO) WHO. Tuberculosis global facts 2019 [05/20/2020. Available from: https://www.who.int/tb/publications/factsheet_global.pdf.
[23] Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. Mamm Genome 2012;23(9-10):713–8.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] Threadgill DW, Churchill GA. Ten years of the collaborative cross. Genetics 2012;190(2):291–4.
[30] 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):e00097. 20.
[31] Hunter RL, Actor JK, Hwang SA, 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.
[32] 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 Intensive Care Med 2005;20(6):354–9.
[33] Major S, Turner J, Beamer G. Tuberculosis in CBA/J mice. Vet Pathol 2013;50(6):1016–21.
[34] 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.
[35] Kus P, Gurcan MN, Beamer G. Automatic detection of granuloma necrosis in pulmonary tuberculosis using a two-phase algorithm: 2D-TB. Microorganisms 2019;7(12):661.

[36] Kramnik I, Beamer G. Mouse models of human TB pathology: roles in the analysis of necrosis and the development of host-directed therapies. Semin Immunopathol 2016;38(2):221–37.
[37] 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.
[38] Harrison DE, Astle CM, Niazi MKK, 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.
[39] 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.
[40] 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.
[41] 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.
[42] Nandi B, Behar SM. Regulation of neutrophils by interferon-gamma limits lung inflammation during tuberculosis infection. J Exp Med 2011;208(11):2251–62.
[43] Smith CM, Proulx MK, Olive AJ, Laddy D, Mishra BB, Moss C, et al. Tuberculosis susceptibility and vaccine protection are independently controlled by host genotype. mBio 2016;7(5):e01516. 16.
[44] Beamer GL, Turner J. Murine models of susceptibility to tuberculosis. Arch Immunol Ther Exp (Warsz) 2005;53(6):469–83.
[45] Niazi MKK, Beamer G, Gurcan MN. A computational framework to detect normal and tuberculosis infected lung from H and E-stained whole slide images. Med Imaging 2017 Digit Pathol 2017;10140:101400J.
[46] 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. Med Imaging Digit Pathol 2018;10581:1058108.
[47] Tavolara TE, Niazi MKK, Beamer G, Gurcan MN. Segmentation of mycobacterium tuberculosis bacilli clusters from acid-fast stained lung biopsies: a deep learning approach. Med Imaging Digit Pathol 2020;11320:113200E.
[48] Tavolara TE, Niazi MKK, Ginese M, Piedra-Mora C, Gatti DM, Beamer G, et al. Automatic discovery of clinically interpretable imaging biomarkers for mycobacterium tuberculosis supersusceptibility using deep learning. EBioMedicine 2020;60:1–100.
[49] Abdulaal A, Patel A, Charani E, Denny S, Alqahtani SA, Davies GW, et al. Comparison of deep learning with regression analysis in creating predictive models for SARS-CoV-2 outcomes. BMC Med Inform Decis Mak 2020;20(1):1–11.
[50] Zhang Z, Pan Q, Ge H, Xing L, Hong Y, Chen P. Deep learning-based clustering robustly identified two classes of sepsis with both prognostic and predictive values. EBioMedicine 2020;62:103081.
[51] Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE – an interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med 2021;113:102036.
[52] Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Sci Rep 2020;10(1):1–11.
[53] Dolezal JM, Trzcinska A, Liao CY, Kochanny S, Blair E, Agrawal N, et al. Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Mod Pathol 2020.
[54] Xu J, Shi J, Cui X, Cui Y, Li JJ, Goel A, et al. Cellular Heterogeneity-Adjusted cLonal Methylation (CHALM) improves prediction of gene expression. Nat Commun 2021;12(1):400.
[55] Chen T., He T., Benesty M., Khotilovich V., Tang Y. Xgboost: extreme gradient boosting. R package version 04-2. 2015:1-4.
[56] Xu Z, Huang G, Weinberger KQ, Zheng AX. Gradient boosted feature selection. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York, New York, USA. Association for Computing Machinery; 2014. p. 522–31.
[57] 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.
[58] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the international conference on machine learning; 2018. p. 2127–36.
[59] Sudharshan P, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P. Multiple instance learning for histopathological breast cancer image classification. Expert Syst Appl 2019;117:103–11.
[60] Multiple instance learning with center embeddings for histopathology classification. In: Chikontwe P, Kim M, Nam SJ, Go H, Park SH, editors. Proceedings of the International conference on medical image computing and computer-assisted intervention. Springer; 2020.
[61] Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data efficient and weakly supervised computational pathology on whole slide images. Nat Biomed Eng 2021:1–16.
[62] Lippi M, Gianotti S, Fama A, Casali M, Barbolini E, Ferrari A, et al. Texture analysis and multiple-instance learning for the classification of malignant lymphomas. Comput Methods Progr Biomed 2020;185:105153.
[63] Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 2019;1(5):236–45.
[64] Deep residual learning for image recognition. In: He K, Zhang X, Ren S, Sun J, editors. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
[65] Robust signal generation and analysis of rat embryonic heart rate in vitro using laplacian eigenmaps and empirical mode decomposition. In: Niazi MKK, Ibrahim MT, Nilsson MF, Sköld A-C, Guan L, Nyström I, editors. Proceedings of the international conference on computer analysis of images and patterns. Springer; 2011.
[66] Cheng Y. Mean shift, mode seeking and clustering. IEEE Trans Pattern Anal Mach Intell 1995;17(8):790–9.
[67] Kumar GSS, Venugopal AK, Kashyap MK, Raju R, Marimuthu A, Palapetta SM, et al. Gene expression profiling of tuberculous meningitis co-infected with HIV. J Proteom Bioinform 2012;5(9):235.
[68] Han JH, Lee S, Park YS, Park JS, Kim KY, Lim JS, et al. IFITM6 expression is increased in macrophages of tumor-bearing mice. Oncol Rep 2011;25(2):531–6.
[69] Yánez DC, Ross S, Crompton T. The IFITM protein family in adaptive immunity. Immunology 2020;159(4):365–72.
[70] Leong NNK, Brombacher F, Dalpke AH, Weitnauer M. Crosstalk between glucocorticoids and IL-4 modulates Ym1 expression in alternatively activated myeloid cells. Immunobiology 2017;222(5):759–67.
[71] Wasser B, Pramanik G, Hess M, Klein M, Luessi F, Dornmair K, et al. Increase of alternatively activated antigen presenting cells in active experimental autoimmune encephalomyelitis. J Neuroimmune Pharmacol 2016;11(4):721–32.
[72] Khaw YM, Cunningham C, Tierney A, Sivaguru M, Inoue M. Neutrophil-selective deletion of Cxcr2 protects against CNS neurodegeneration in a mouse model of multiple sclerosis. JNeuroinflamm 2020;17(1):1–12.
[73] Dorhoi A, Dorhoi SH. Pathology and immune reactivity: understanding multidimensionality in pulmonary tuberculosis. Semin Immunopathol 2016;38(2):153–66.