Original Article

Improving Generalization of Deep Learning Models for Diagnostic Pathology by Increasing Variability in Training Data: Experiments on Osteosarcoma Subtypes

Haiming Tang[1], Nanfei Sun[2], Steven Shen[3]

[1]Department of Pathology and Laboratory Medicine, Yale New Haven Hospital, New Haven, Connecticut, USA,
[2]Department of Management Information System, College of Business, University of Houston Clear Lake, Houston, Texas, USA,
[3]Department of Pathology and Genomic Medicine, Houston Methodist Hospital, Houston, Texas, USA

Submitted: 15-Sep-2020 Revised: 19-Apr-2021 Accepted: 29-Apr-2021 Published: 04-Aug-2021

Abstract

Background: Artificial intelligence has an emerging progress in diagnostic pathology. A large number of studies of applying deep learning models to histopathological images have been published in recent years. While many studies claim high accuracies, they may fall into the pitfalls of overfitting and lack of generalization due to the high variability of the histopathological images. **Aims and Objects:** Use the model training of osteosarcoma as an example to illustrate the pitfalls of overfitting and how the addition of model input variability can help improve model performance. **Materials and Methods:** We use the publicly available osteosarcoma dataset to retrain a previously published classification model for osteosarcoma. We partition the same set of images into the training and testing datasets differently than the original study: the test dataset consists of images from one patient while the training dataset consists images of all other patients. We also show the influence of training data variability on model performance by collecting a minimal dataset of 10 osteosarcoma subtypes as well as benign tissues and benign bone tumors of differentiation. **Results:** The performance of the re-trained model on the test set using the new partition schema declines dramatically, indicating a lack of model generalization and overfitting. We show the additions of more and moresubtypes into the training data step by step under the same model schema yield a series of coherent models with increasing performances. **Conclusions:** In conclusion, we bring forward data preprocessing and collection tactics for histopathological images of high variability to avoid the pitfalls of overfitting and build deep learning models of higher generalization abilities.

Keywords: Artificial intelligence, computer vision, deep learning, diagnostic pathology, osteosarcoma, overfitting

Introduction

Artificial intelligence (AI) has been successfully applied to many tasks including image detection and classification, sound processing, and natural language processing. In the area of image classification and detection of everyday objects, AI algorithms such as deep learning have been tremendously successful. Nowadays, smartphone cameras can detect user faces and surrounding objects accurately.

There has been an exponential growth in the application of AI in the health and medical fields. Deep learning algorithms have been built to segmentation organs from X-rays, computed tomography, and magnetic resonance imaging images. Recently, physicians and computer scientists have jointed work in building deep learning algorithms to diagnose COVID-19 pneumonia features.[1] Up to September 2020, the Food and Drug Administration has approved 76 AI algorithms in the field of diagnostic radiology, according to the Data Science Institute of American College of Radiology.[2]

In the field of diagnostic pathology, we have also seen deep learning applications in histopathological images. Many of

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_78_20

Address for correspondence: Dr. Haiming Tang,
Yale New Haven Hospital, 20 York Street, Ste East Pavilion 2-610,
New Haven, CT, 06510, USA.
E-mail: ningzhithm@gmail.com

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Tang H, Sun N, Shen S. Improving generalization of deep learning models for diagnostic pathology by increasing variability in training data: Experiments on osteosarcoma subtypes. J Pathol Inform 2021;12:30.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/30/323178

these applications are specifically aimed at hematoxylin and eosin (H and E)-stained images and have the potential of transforming diagnostic pathology, like what has already been happening in the field of radiology.[3] Progresses have been made in areas such as prostate cancer biopsy diagnosis[4] and evaluation of cervical cytology.[5] Research group in University of Pittsburg Medical Center claimed that they have developed a deep learning model for prostate cancer surveillance with high performance and have deployed the model in routine work of Maccabi Healthcare Services in Israel.[6]

There is no doubt that AI has tremendous power and has great potential in application in diagnostic pathology. With the emerging progress of AI in pathology images, there will be great progress in the near future. However, deep learning models are not without challenges: "Model generalization" is the most common hurdle. As described in book Deep Learning,[7] "The central challenge in machine learning is that we must perform well on new, previously unseen inputs – not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called generalization."

The deep learning model development routine is to split the input data to training and testing datasets. The model is developed using the training dataset and tested on the testing dataset. Model generalization abilities are estimated on the testing dataset or additional validation dataset. Cross-validation techniques are often used for assessing model generalization.

The most common pitfall of a deep learning model is overfitting. Overfitting means the model goes through too much learning and the model performs well on the training dataset, but poorly on the new data. This arises from the lack of variability between the test or validation dataset and the data in the real world.[8]

A large number of studies of applying deep learning models to histopathological images have been published in recent years, and many of these studies have a very similar schema: the authors collect sets of images of different categories, such as normal tissue, benign tumor, malignant tumor, different stages of tumors, or metastatic tumor. The images are randomly split into train and test datasets. Deep learning model is trained on the training dataset and model performance is reported from the test dataset. The authors usually claimed high model performances in terms of high accuracy and high sensitivity and specificity. Model accuracies were often claimed to be higher than 0.99. Many studies even claimed that the models have higher performance than experienced pathologists in the specific test sets.

However, it is likely that some of these studies fall into the pitfall of overfitting. The models surely perform very well on the specific test sets used in the author's study, but because the training and testing datasets lack variability compared with the real-world data, the trained models will fail on the new real-world data that is of the same diagnosis, but with different image presentations.

Here, we looked at the performance and generalization of a deep learning model in a previously published paper.[9] The authors used osteosarcoma biopsy images to build a classification model for benign tissue, viable tumor, and nonviable tumor. We rebuilt the deep learning model using the same image datasets and the same model schema that the authors have made publicly available at the Cancer Imaging Archive (TCIA).[10,11] We achieved comparable model performance using the same train/test schema as stated in the paper. We then use the same dataset but a different train/test split schema for a new model: all images that come from one specific patient were used as the test dataset and the images of all other patients were used as training set. This refitted new model using the same deep learning schema shows good performance on the training dataset but poor performance on the test set of one patient data left out, indicating a possibility of overfitting. We suspect that the overfitting comes from the similarity among images of the same patient. The overfitting problem is exposed after we restrict the test dataset to patient images that are not included in the training dataset. In other words, the new model is not generalizable to the one patient left out.

Histopathological images are notoriously highly variable. Even experienced pathologists sometimes do not have consensus diagnosis. The variation comes from many levels, such as specimen preparation and artifacts, patient level variations, tumor stages, tumor types/subtypes, and tumor heterogeneity. Some tumor types are especially highly variable. Osteosarcoma, for example, has around 10 different morphologic subtypes.[12] Some subtypes are very similar in morphology to benign bone tumors such as osteoma and osteoid osteoma.[13] It is our hypothesis that the lack of variability in the training data can be a main obstacle in building a robust diagnosing model.

To illustrate the effects of lacking variability in deep learning models, we built a series of deep learning models for classifying osteosarcoma vs. benign tissue or benign bone tumors using different combinations of training datasets. We collected histopathological images for each of the osteosarcoma subtypes, benign bone tumors that should be differentiated with osteosarcoma including osteoma and osteoid osteoma, as well as benign tissues that may appear in bone biopsies, such as normal bone, soft bone, muscle, and connective tissues. The test dataset we built is fixed for all models; it is composed of all subtypes of osteosarcoma and all types of benign tissues and tumors. In contrast, for the training datasets, each of the subtypes of osteosarcoma and add-up of different subtypes are used. We find that while the model performances on the training datasets are consistently high, the performances on the fixed test set composed of all osteosarcoma subtypes increase as more and more subtypes are included in the training dataset. From this experiment, our primitive conclusion is that higher variability in training dataset is beneficial for a robust model to be applicable to the real-world data. Tumor subtype classification, in a way, is the human intelligence in clustering the tumor based on their variability. The hypothesis that the inclusion of multiple tumor subtypes is one of the most efficient ways to boost the data variability; thus, improving the robustness of the models has also been observed and supported in other studies.[14]

Thus, using the example of osteosarcoma subtypes, we demonstrated the effects of lacking variability in training data on the model generalization ability. We also proposed a methodology to check the issue of overfitting of the deep learning models. Our study proposes a possible reference for the development of highly robust deep learning models for diagnostic pathology in the future.

Methods

The Cancer Imaging Archive dataset

Leavy *et al*. have made the osteosarcoma data they used to train the classification model publicly available. We downloaded the data from TCIA website.[10,11]

The dataset is composed of 1144 H and E-stained osteosarcoma histology images, from the four patients who had been treated at Children's Medical Center, Dallas, between 1995 and 2015. Table 1 shows the crosstab table of different tumor types and different patients.

Out of these images, there are 536 (47%) nontumor images, 263 (23%) necrotic tumor images, and 292 (25%) viable tumor tiles. A total of 53 images have unclear status between viable and nonviable; they were emitted in model fitting.

The images were of 1024 × 1024 pixels each; they were split into 128 × 128 image tiles. The same data preprocessing steps including red, green, and blue channel to Lab color space conversion, addition of original 1024 × 1024 images to training data, and removal of images containing only white pixels or empty background pixels were followed.

Data augmentation including vertical and horizontal flip and height and width shift was applied using Keras ImageDataGenerator class.

We tried two different methods to generate the train and test datasets. The first method was the conventional random split of the whole dataset by 0.7/0.3 ratio similar to the method used in the original journal. As the entire dataset is composed of images from four patients, we suspected that there is lack of variability in the training dataset and the high performance reported by the authors probably comes from overfitting. To illustrate that, in the second method, we used all the images of patient "Case 4" as the test dataset, while the images of the rest three patients were used as the training set. The patient "Case 4" of this dataset contained all 3 types of the tumor, thus making it a good test set.

Table 1: Crosstab table of different tumor types and different patients for the cancer imaging archive osteosarcoma dataset

Tumor_type	Patient_id				Total
	P9	Case 3	Case4	Case48	
Nontumor	212	110	78	136	536
Nonviable-tumor	0	171	90	2	263
Viable	0	3	87	202	292
Viable: Nonviable	0	1	22	30	53
Total	212	285	277	370	1144

All subtypes dataset and benign dataset

Osteosarcoma subtypes include conventional variants, surface types, and other variants like small cell, extraskeletal, and secondary osteosarcoma like complicating Paget's disease. Conventional variants include osteoblastic, chondroblastic, telangiectatic, and fibroblastic subtypes,[15] and surface osteosarcomas include periosteal and parosteal and high-grade surface types.[16] In Figure 1, we show the different subtypes of osteosarcoma.

To maximize the diversity in the training data for diagnosing osteosarcoma, we collected histopathological images of osteosarcoma by subtypes. As the aim of this study is to illustrate the effects of data variability on model performance instead of building a robust classification model for osteosarcoma, minimal numbers of histopathological images were collected for ease of training.

Images were collected from online sources and reviewed by experienced pathologists. Osteoblastic (41.7%) and chondroblastic (20.8%) subtypes were reported to be the more common subtype; thus, we collected relative more images for these subtypes. However, we do not claim that the composition of the images reflects the ensemble osteosarcoma in the real world due to its complex nature. Its effects are covered in the discussion part.

The design of the benign dataset is also aimed to maximize the variability, but within a reasonable range. We collected the benign tissue types that commonly appear in bone biopsy, including bone, soft bone, muscle, tendon, and connective tissue proper.

We also collected histopathological images of the benign bone tumors that should commonly be differentiated with osteosarcoma, including osteoma and osteoid osteoma. Sample images of the benign dataset are included in Figure 2.

The images in the subtype and benign datasets were of various sizes. Image tiles of size 128 × 128 pixels were extracted starting from the top left corner, toward the right and bottom edges. Tiles that cross the right and bottom edges were discarded. The images were rotated 90°, 180°, and 270°, and image tiles were collected to prevent the learning of position-dependent features by models.

The collected images and their sources are available at https://github.com/haimingt/osteosarcoma_subtype_modeling/tree/master/subtypes.

Model training and evaluation

Keras implementation of the same convolution neural network

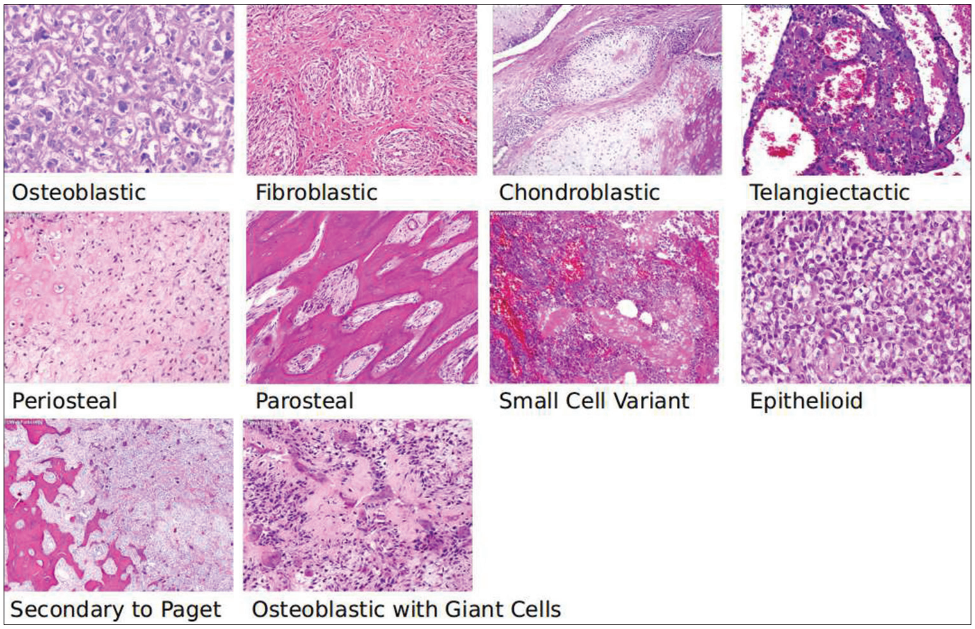

Figure 1: Sample images of different osteosarcoma subtypes we have collected for Experiments E and F showed in Table 2

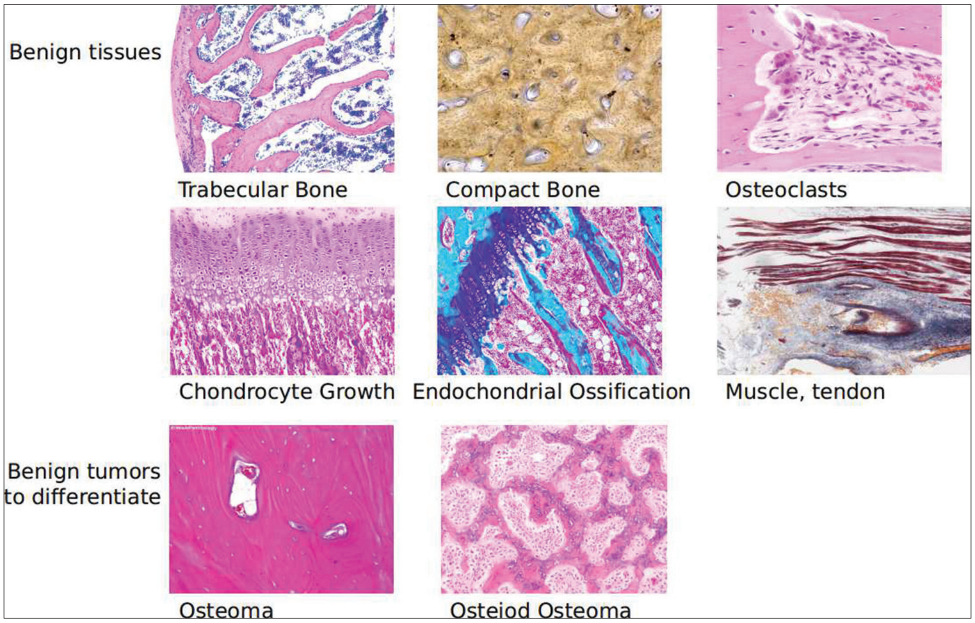

Figure 2: Benign dataset of Experiment E and F consisting of benign bone tissues and 2 types of benign tumors, osteoma, and osteoid osteoma

schema was used for all experiments. While there are numerous variations of schema, we used the schema that was included in the previous published paper, which was extended from the classic LeNet5 schema by adding more convolution layers. The model details can be found in Supplemental Materials. Each of the training processes consisted of 25 epochs.

A series of experiments were performed in our study [Table 2].

Experiment A and B used the TCIA datasets to train and test models. Experimental C and D applied the models in A and B to the test set composed of all osteosarcoma subtypes, benign tissues, osteoma, and osteoid osteoma. Experiment E used just 1 type of osteosarcoma to train but predict data that contains all osteosarcoma subtypes. Experiment F used combinations of the different subtypes in the training data in an add-up manner; for each step, an additional subtype was added to the training data, while the test dataset was the same as compared to experiment F. The order of the "step-by-step add-up" followed the rankings of the performance of each tumor subtype as in Experimental E, the subtypes with smaller area under curve (AUC) were added first to the combination models.

The metrics used for evaluating model performance include the AUC, accuracy, precision, and recall.

Table 2: Experiments set up and performance summary

Experiment label	Training set	Test set	Specific subtype of osteosarcoma	AUC
A	70% of TCIA set	30% of TCIA set	NA	0.831406
B	TCIA set (P9, case 3, case 48)	TCIA set (Case 4)	NA	0.700612
C	70% of TCIA set	30% of all subtypes of osteosarcoma+all benign tissues , osteoma and osteoid osteoma	NA	0.55
D	TCIA set (P9, Case 3, Case 48)		NA	0.29
E	70% of dfifernt combinatiosn of otsteosarcoma subtypes+70% of all benign tissues , osteoma and osteoid osteoma		Smallcellvariant_fibroblastic	0.471908
			Smallcellvariant_fibroblastic_periosteal	0.524636
			Smallcellvariant_fibroblastic_periosteal_telangiectactic	0.528312
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget	0.598128
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget_epithelioid	0.863568
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget_epithelioid_withgiantcells	0.851232
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget_epithelioid_withgiantcells_parosteal	0.840264
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget_epithelioid_withgiantcells_parosteal_ osteoblastic	0.885664
			Smallcellvariant_fibroblastic_periosteal_telangiectactic_ complicatingpaget_epithelioid_withgiantcells_parosteal_ osteoblastic_chondroblastic	0.88704
F	70% of one specific type of osteosarcoma+70% of all benign tissues , osteoma and osteoid osteoma		Chondroblastic	0.663656
			Complicatingpaget	0.451864
			Epithelioid	0.457356
			Fibroblastic	0.392112
			Osteoblastic	0.630856
			Parosteal	0.542764
			Periosteal	0.396016
			Smallcellvariant	0.388852
			Telangiectactic	0.399856
			Withgiantcells	0.457676
G	70% of one specific type of osteosarcoma+70% of all benign tissues	30% of all subtypes of osteosarcoma+all benign tissues	Chondroblastic	0.916284
			Complicatingpaget	0.826564
			Epithelioid	0.906808
			Fibroblastic	0.732032
			Osteoblastic	0.94616
			Parosteal	0.761204
			Periosteal	0.783428
			Smallcellvariant	0.535136
			Telangiectactic	0.569512
			Withgiantcells	0.660888

TCIA: The cancer imaging archive, NA: Not available, AUC: Analytical ultracentrifugation

To ensure comparability, we manually converted the predictions of Experiments C and D from 3 categories (benign, noviable and viable) to 2 categories (non-osteosarcoma and osteosarcoma) by combining prediction of benign and viable as non-osteosarcoma. The predicted probability of osteosarcoma was set to be equal to the probability of nonviable tumor, while the probability of benign tissue or tumor was set to be equal to the probability of benign and viable tumors combined.

Results

Overfitting caused by lack of variability in the Cancer Imaging Archive training data

In Experiment A, we followed the routine way of splitting data into training and testing data. The model performance is good on the testing data, with AUC = 0.83. However, in Experiment B, while the training data were from the first three patients, and the model was tested on the 4th patient, the performance dropped to an average AUC of 0.70. It proves that the model trained on the three patients cannot predict the 4th patient well.

In addition, as shown in Figure 3, the training process in Experiment A shows a stable and consistent pattern between the training and testing sets. The performance of Experiment B shows discordant trends between train and test sets; the training set has an upward trend through epochs, whereas the test set has a downward trend. It means that the more training

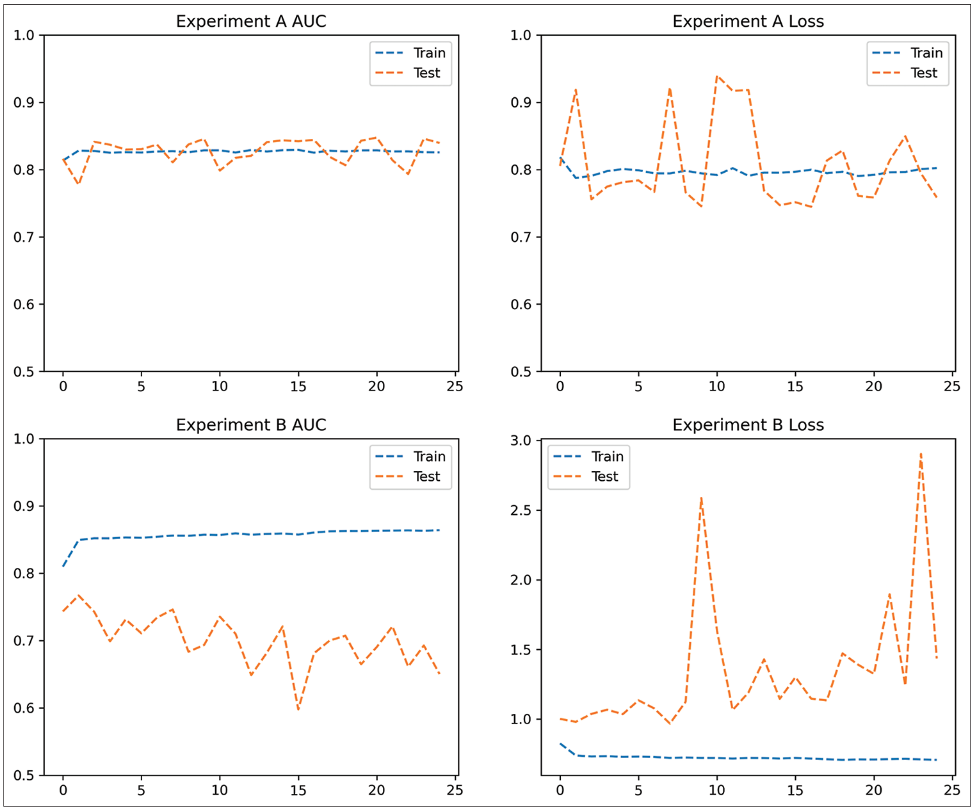

Figure 3: Model metrics: Loss and area under curve during training epochs, upper Experiment A, lower Experiment B

is performed and the better the performance of the model on the training set, the worse the performance is on the test set. This lack of the model generalization on the new data reveals the typical error of model overfitting on the training data.

We then applied the models in Experiment A and B to the newly collected test sets that consist of 10 different subtypes of osteosarcoma, benign tissues, and 2 benign bone tumors. As examined by the pathologist, the osteosarcoma images in the TCIA datasets all belong to the osteoblastic type. Thus, the test dataset contains far more variability than the training data used in Experiment A and B. It is not surprising that the models of A and B will lack generalization toward these new data. Moreover, the performance confirms our hypothesis, the model in Experimental A has an AUC of 0.57, and the model in Experiment B has an AUC of 0.40 [Figure 4].

Overfitting caused by only one subtype of osteosarcoma in training data

In Experiment E, training data contain only one subtype of osteosarcoma, all the benign tissues and benign bone tumors, while the testing data contain all 10 subtypes. We designed this experiment to roughly represent the situation in real life, when the training data only reflect a very small part of the complexities of the real-world data. Thus, we expect the models in Experiment E to perform badly in the overfitting way.

As shown in Figure 3, this issue is well illustrated by the case of using chondroblastic subtype in the training dataset. The performance on the training data improves epoch by epoch, showing better model fitting upon each step. However, the performance on the test dataset shows large fluctuations, indicating that the features learned by the training data are not the "correct" features to differentiate the osteosarcoma versus nonosteosarcoma bone tissues. Figure 5 shows the metrics of the model using only the chondroblastic subtype to train in Experiment E. Plots for models using other subtypes can be found in Supplemental Materials.

Figure 6 summarizes the performances of Experiment E for each of the subtypes used. It shows the boxplot of the AUC of the 25 epochs for each of the models using only one subtype. As expected, the general performances of most models are unsatisfactory, with average AUC <0.7. Moreover, for models using parosteal, osteoblastic, and chondroblastic, the performances are slightly better, but showing great fluctuations, indicating a lack of fit of the trained model.

Addition of more subtypes in training data increases model performance

We then ranked the performance of the models of each subtype from low to high by the average AUC of the 25 epochs and sequentially added one more subtype to the training data. The models using different training sets were then tested on the same dataset that was used in Experiment E.

Figure 7 shows the boxplot of the AUC of the 25 epochs of subtype models in Experiment E on the same test dataset. It shows a clear pattern of improved model performance with each step of adding more osteosarcoma subtypes to the training set. Specifically, the AUC is 0.39 for the model of using small cell variant subtype only, AUC performance increases to 0.47 when a combination of small cell variant and fibroblastic subtype was used, and the AUC for the final model using all subtypes is 0.89.

The only violation of the trend of performance increase comes from the addition of parosteal subtype, which reduced the model AUC performance slightly from 0.85 to 0.84. The exact cause of the decline is unclear. Our suspect is that parosteal subtype arises from the bone surface and it is generally well differentiated of a lower stage (Stage I and II).[16] In the images we have selected, there are several images with areas of chondroid differentiation. The cartilage is present at the periphery of the lesion and may resemble a benign cartilage tissue. We suspect that the addition of this subtype, although adding slightly more images and variability to the train model, may not overcome the error in the test dataset caused by similarities between the chondroid differentiation in the parosteal subtype and the normal cartilage tissues.

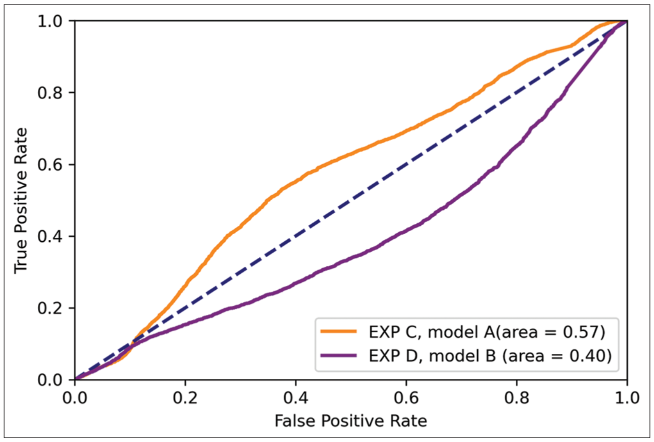

Figure 4: Receiver operating curves for Experiments C and D, which are the performances of the models in Experiment A and B applied to the test set composed of all subtypes of osteosarcoma, benign tissues, and benign bone tumors

From this experiment, we can conclude that to generate a deep learning model of generalization ability in the field of histopathology, the training dataset should contain enough data that is diverse enough to cover all kinds of images the model will be applied to. In the case of developing a deep learning model for diagnosing osteosarcoma versus nonosteosarcoma, including all subtypes of osteosarcoma may be a good method to increase model variability and robustness.

Discussions

It is important to note that our experiments only use minimal data. The purpose of this article is to perform experiments that illustrate the issues of overfitting and the lack of generalization in the development of deep learning models for histopathology.

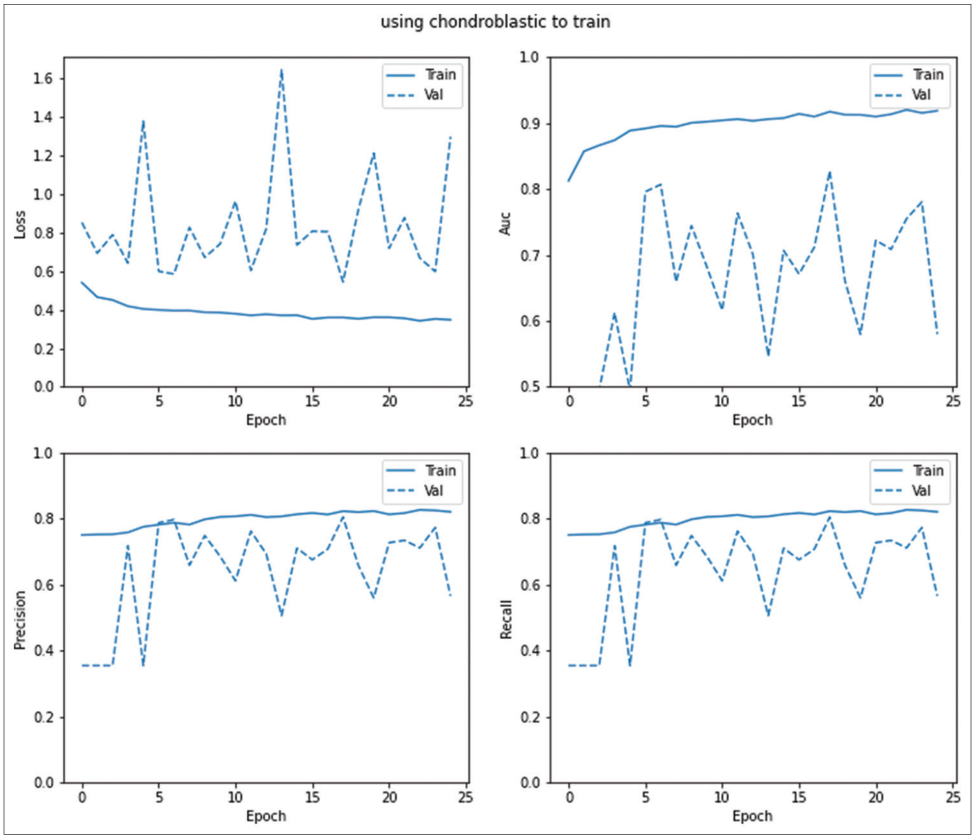

Figure 5: Metrics of the model using only chondroblastic subtype to train in Experiment E

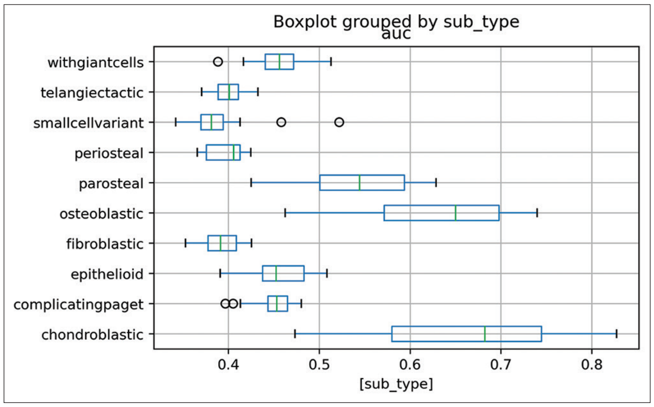

Figure 6: Boxplot of the area under curve of the 25 epochs of subtype models in Experiment E on the same test dataset

The test data we collected have more variability than data containing only limited subtypes of sarcoma, but the data variety in the real world is significantly greater than what we have included in this study. The models we developed are by no means production models that can be applied to real-world osteosarcoma data. However, we have proposed a framework regarding if we have a large dataset of various cancer subtypes, how to build a model that is more robust to the varieties of the images, and how to avoid overfitting during the model building process. By separating the training images based on different subtypes of corresponding diseases, this framework allows users to build a series of coherent models, and based on the performances of these models, users can thus produce a performance curve of these models. Ideally, the performance curve will grow nonlinearly until it reaches its upper limit caused by the law of diminishing marginal utility. We can derive similar performance curves using different numbers and qualities of training images as well. By changing the training image sizes, qualities, and varieties, we can potentially assess the robustness, reliability, and confidence of these models and finally derive a confident score of the model for diagnosing purposes.

The routine schema of deep learning model building as that used in Experiment A may create the issue of overfitting. Many researches collect a limited number of images from a small number of patients. Moreover, these images are split into training and test datasets randomly.

Because of the intrinsic similarities among the images from the same patient, there will be overly exaggerated similarity between the training and testing datasets. Thus, the high performance converged by the test dataset is commonly overestimated and the model is commonly overfitted.

To spot this issue, we recommend building the test dataset in a more careful way that excludes similar data from the training dataset. Researchers can use the images from 1 or more new patients, while the model is trained on images from other patients. Another recommendation is not to split the image tiles or patches from the same large image into train and test dataset, as image tiles or patches may share great similarities and can affect the model evaluation.

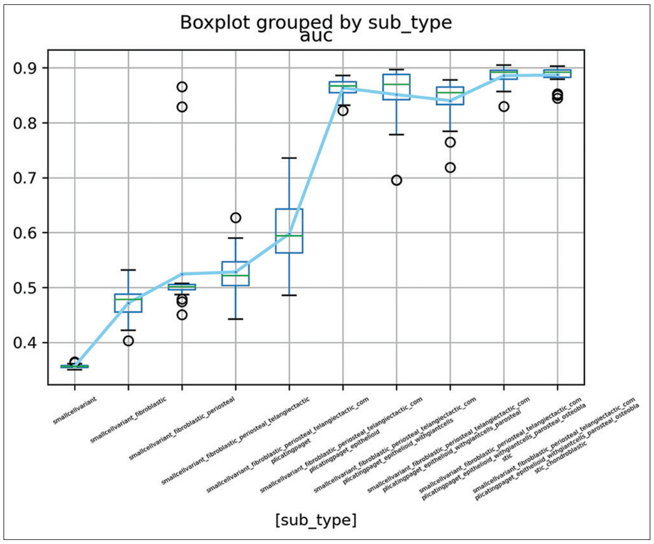

Figure 7: Boxplots of the area under curve of the 25 epochs of models that add up different subtypes in Experiment F

Histopathological images are of great variability. Research has shown the high interobserver variability with regard to histological grade of differentiated tumours.[17] Review paper summarizes that the diagnostic variability in breast cancer could be attributed to three overall root causes: (i) pathologist related, (ii) diagnostic coding/study methodology related, and (iii) specimen related. Most pathologist-related root causes were attributable to professional differences in pathologists' opinions about whether the diagnostic criteria for a specific diagnosis were met, most frequently in cases of atypia.[18]

Experiments E and F show that the lacking of variability in the training data greatly affects the model performance. When more training data reflects more complexity of the real-world data, model performance increases.

A dataset should be built by maximizing the reasonable variability. The "do more less well" principle[19] for sampling efficiency of stereological studies in biology indicates that the variations at the lower levels of sampling are of minor importance. In particular, the "biological variation" between different subjects plays an all-important role, whereas the feature-to-feature variation on sections is of negligible importance. The database by the original authors is generally large collection 1144 histology images. However, these images were from only four patients. Thus the database is still lacking in variability. In contrast, if the database contains the bone histology images from around 1000 patients, the spectrum of osteosarcoma and normal bone tissue appearances can be much better represented.

Collecting samples from different subjects is a key component for data variability. In addition to the subject level variability, it may be easiest to the increase data variability by incorporating pathological knowledge grounds. For example, the inclusion of different stages or types of a cancer. Like in the case of ostesarcoma diagnosing model, all different subtypes of osteosarcoma should be incorporated to increase the data variability.

This, however, will uninhibitedly require a large number of images of various types from a large pool of patients. It is often noted that due to the patient confidentiality, the histopathological images used in many published studies were not publicly accessible. Recently, more digital pathology datasets such as CAMELYON[20] have become publicly available and have pushed the frontiers of deep learning in pathology informatics. With the growing abundance of data availability and variability, we will be able to build more robust deep learning models for computer-aided diagnosis systems.

Conclusion

In this article, we examined the pitfalls of overfitting and the lack of generalization in deep learning models in histopathological images through a series of experiments on osteosarcoma. We demonstrated that the lack of variability in the training data can lead to overfitting of the models and the random split of the train and test dataset from the same patient or images may disguise the overfitting problem. We also showed that with the addition of more data with increased variability to the training data, models can achieve higher levels of robustness. From these, we bring forward data preprocessing and collection tactics to build deep learning models of higher generalization abilities by avoiding the pitfalls of overfitting.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, *et al.* Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020;181:1423-33.e11.
2. American College of Radiology Data Science Institute. Available from: https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms. [Last accessed on 2020 Sep].
3. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
4. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. Lancet Oncol 2020;21:233-41.
5. Wentzensen N, Lahrmann B, Clarke MA, Kinney W, Tokugawa D, Poitras N, *et al.* Accuracy and efficiency of deep-learning-based automation of dual stain cytology in cervical cancer screening. J Natl Cancer Inst 2021;113:72-9.
6. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, *et al.* An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: A blinded clinical validation and deployment study. Lancet Digit Health 2020;2:e407-16.
7. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Massachusetts: MIT Press; 2016. p. 110.
8. Rice L, Wong E, Kolter JZ. Overfitting in adversarially robust deep learning. arXiv 2020.
9. Mishra R, Daescu O, Leavey P, Rakheja D, Sengupta A. Convolutional neural network for histopathological analysis of osteosarcoma. J Comput Biol 2018;25:313-25.
10. Leavey P, Sengupta A, Rakheja D, Daescu O, Arunachalam HB, Mishra R. Osteosarcoma data from UT Southwestern/UT Dallas for viable and necrotic tumor assessment [Data set]. Cancer Imaging Arch 2019. [doi: 10.7937/tcia. 2019.bvhjhdas].
11. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, *et al.* The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging 2013;26:1045-57.
12. Lindsey BA, Markel JE, Kleinerman ES. Osteosarcoma overview. Rheumatol Ther 2017;4:25-43.
13. Bukhari MH. In: Qamar S, editor. Differential Diagnosis of Osteogenic Tumors in the Context of Osteosarcoma. Ch. 2. Rijeka: IntechOpen; 2019. DOI: 10.5772/intechopen.85190.
14. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683.
15. Durfee RA, Mohammed M, Luu HH. Review of osteosarcoma and current management. Rheumatol Ther 2016;3:221-43.
16. Kumar VS, Barwar N, Khan SA. Surface osteosarcomas: Diagnosis, treatment and outcome. Indian J Orthop 2014;48:255-61.
17. Komuta K, Batts K, Jessurun J, Snover D, Garcia-Aguilar J, Rothenberger D, *et al.* Interobserver variability in the pathological assessment of malignant colorectal polyps. Br J Surg 2004;91:1479-84.
18. Allison KH, Reisch LM, Carney PA, Weaver DL, Schnitt SJ, O'Malley FP, *et al.* Understanding diagnostic variability in breast pathology: Lessons learned from an expert consensus review panel. Histopathology 2014;65:240-51.
19. Gundersen HJ, Osterby R. Optimizing sampling efficiency of stereological studies in biology: Or 'do more less well!'. J Microsc 1981;121:65-73.
20. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. Gigascience 2018;7:giy065.

Supplemental Materials

Supplemental materials and our histopathology image collection of osteosarcoma subtypes as well as benign tissues and benign bone tumors can be found at https://github.com/haimingt/osteosarcoma_subtype_modeling