Biased Data, Biased AI: Deep Networks Predict the Acquisition Site of TCGA Images

Taher Dehkharghanian

McMaster University

Azam Asilian Bidgoli

Ontario Tech University

Abtin Riasatian

University of Waterloo

Pooria Mazaheri

Ontario Tech University

Clinton J.V. Campbell

McMaster University

Liron Pantanowitz

University of Michigan

H. R. Tizhoosh

University of Waterloo

Shahryar Rahnamayan (✉ shahryar.rahnamayan@ontariotechu.ca)

Ontario Tech University

Research Article

Keywords: Cancer Genome Atlas (TCGA), digital pathology, histopathology

Posted Date: October 4th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-943804/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Deep learning models applied to healthcare applications including digital pathology have been increasing their scope and importance in recent years. Many of these models have been trained on The Cancer Genome Atlas (TCGA) atlas of digital images, or use it as a validation source. This study shows that there are tissue source site (tss) specific patterns of TCGA images that could be used to identify contributing institutions without any explicit training. Furthermore, it was observed that a model trained for cancer subtype classification has discovered such tss specific patterns within digital slides to classify cancer types. Digital scanner configuration and noise, tissue stain variation and artifacts, and source site patient demographics are among factors that likely account for the observed bias. Therefore, researchers should be cautious of such bias when using histopathology datasets for developing and training deep networks.

Introduction

The Cancer Genome Atlas (TCGA) dataset is one of the largest publicly available digital pathology repositories.¹ The TCGA repository contains histopathology whole slide images (WSIs) of 32 cancer subtypes as well as their associated metadata (e.g., clinical information and reports).² Many research groups have trained and validated deep learning models using TCGA images for various tasks ranging from cancer type classification,^{3–8} tissue segmentation^{9,10} and somatic mutation prediction,^{3,5,8,11} to survival estimation.^{4,12}

One crucial factor that seems to have been widely ignored is the internal bias that originates from the institutions that contributed WSIs to the TCGA dataset. Slides provided by different hospitals may have unique characteristics (e.g., variation in tissue processing, tissue stain protocols, stain quality, color intensity) and other confounding factors including being scanned by specific whole slide scanning hardware platforms, unique imaging protocols (e.g., image compression), and proprietary image file formats. Furthermore, local demographics may lead to batch bias in any given hospital or clinic. The existence of these hospital-identifying factors incorporated into image datasets with accompanying metadata may cause deep models to learn irrelevant visual clues abrogating the generalization of tissue morphology when tested on external validation datasets, i.e., unseen hospitals/clinics.

Image search is a specific application of deep learning in histopathology, where pre-trained deep neural networks are used for feature extraction, i.e., feeding images into a pre-trained model and using the deep feature output of a certain layer for image representation.¹³ Some researchers have utilized networks, pre-trained on images of non-medical objects^{14,15} such as the ImageNet dataset,¹⁶ while other implemented models fine-tuned or trained on pathology images.^{17–19} The underlying idea of image search is that deep features extracted from a query image can be compared to all WSIs in a database in a computationally efficient manner, allowing for diagnostic support by locating and retrieving cases of similar morphology. Unfortunately, such an approach may be subject to bias if the feature extractor is trained on specific institutional datasets and potential “hidden” biases are not accounted for. Logically, this bias affects any other operation such as segmentation, prediction and classification.

Howard et al. showed that the distribution of clinical information in TCGA data such as survival and gene expression patterns significantly differ among samples provided by different institutions.²⁰ The authors report that a deep learning model utilizing image search could be trained employing histopathology images of six cancer types to accurately discriminate submitting institutions. Additionally, the authors found that it is not possible to effectively obfuscate source sites through common stain normalization techniques. These two findings suggest that some models that claim to predict prognosis or mutation states are perhaps not doing anything more than source site detection. In other words, “biased models” are likely identifying the institution that submitted the sample rather than actually doing what they claim to do. Indeed, DeGrave et al. observed that machine learning models trained on radiographic images have a tendency to learn medically irrelevant shortcuts, usually attributable to biases in data acquisition, instead of the actual underlying pathology.²¹

In this report, we used deep features extracted by a deep neural network trained on TCGA WSIs to evaluate biased learning that could lead to distinguishing the source institution that submitted digital slides. This network was trained as a classifier to serve as domain specific feature extractor for image search in histopathology images.

We address the following questions in this paper:

1. How difficult is it to detect tissue source sites? Is it necessary to explicitly train a deep neural network, or can this be done without training? In other words, does a network pre-trained on non-medical images pick up covert hospital-identifying patterns without any further weight optimization?
2. Does a network, trained for a task that is not equally distributed among contributing medical institutions, learn to distinguish these contributing medical institutions? If true, this suggests that deep networks may use irrelevant clues in the data for decision-making which would erode generalization to images from unseen hospitals.
3. Are tissue source site-identifying patterns detectable in all TCGA projects, rather than just within certain cancer types? If they persist in overall data regardless of cancer type, those patterns are most likely histopathologically irrelevant and unrelated to morphologic similarities between samples provided by each institution.

Results

We aimed to assess the effect of bias of contributing institutions in the learning process of networks that had been trained on the TCGA dataset for cancer type classification. Deep feature vectors were used for tissue source classification in order to verify whether such bias affects what deep neural networks learn from training data, or in other words whether some supposedly irrelevant clues in data would mislead the network. Deep features were extracted from WSIs by the KimiaNet proposed by Riasatian et al.¹⁹ It is important to note that the feature extractor network employed had been trained for cancer subtype classification and not for source site detection. On average, to represent tissue samples, KimiaNet uses approximately 55 tissue patches per WSI through clustering the tissue samples in different groups.

KimiaNet was selected because the information about its training and test data is publicly available. It is a deep neural network that borrowed the DenseNet-121 topology and was trained to classify 30 cancer subtypes from tissue patches of size 1000 x 1000 pixels of paraffin-embedded hematoxylin and eosin diagnostic WSIs at 20x magnification. There were 8,579 tissue samples from 141 contributing institutions. The distribution of cases among institutions was imbalanced; 57 institutions provided samples of only one specific cancer type. The real number might be slightly lower than 57 as the tissue source site names were sometimes inconsistent, and we might have mistakenly considered two source sites to be different while they are actually the same. For example, “Emory University” and “Emory University - Winship Cancer Institute” were deemed to be separate institutions.

Tissue patches of size 1000 x 1000 pixels at 20x magnification were sampled from each WSI by the process described by Kalra et al.¹⁵ These tissue patches were fed into both KimiaNet (trained on TCGA images) and DenseNet121 (pre-trained on ImageNet dataset) to extract deep features, i.e., the output of the last pooling layer. These deep feature vectors were of length 1024. The extracted deep feature vectors were later used for tissue source institution identification. This process is depicted in Figure 1.

In order to prevent institutions with a high number of samples from skewing the results, we divided institutions into dominant (those that contributed to more than 1% of WSIs) and the remaining institutions as less-dominant (with fewer WSI contributions). We trained models for source site institution classification separately for these two groups. There were 24 institutions that provided more than 1% of WSIs, which constituted around 74% of all WSIs, while 117 institutions provided 26% of the WSIs.

An artificial neural network with two fully connected hidden layers (500 and 200 neurons, respectively) was trained to classify source site institutions from deep features of tissue patches. The output of this network was adjusted with respect to the number of institutions in each group. WSIs were divided into train and test datasets with approximately a 90 to 10 ratio. WSIs of patients that had more than one WSI were only used for training to prevent any confounding interference that might lead to overestimation of classification accuracy. The process of train-test split and training was repeated 30 times to establish a more reliable estimation of classification performance.

For the models trained on DenseNet’s deep features, the classification accuracy was 70% among the dominating institutions, and 53% for the less-dominant group. The features were extracted by a network that was not trained on pathology images. Nevertheless, this network was able to pick up clues to identify tissue source institutions. For instance, it was plausible to almost perfectly identify the source of samples provided by Indivumed Inc. This finding suggests that there exist some data acquisition-related patterns that are detected by a deep network but that were not specifically utilized for supervised trained on pathology images. For the models trained on KimiaNet’s features, the average accuracy was 86% for the dominating institutions; detailed results are shown in **Supplementary Table 1**. For the less-dominating institutions, the average accuracy was 75% whereas 8 institutions had F1-scores higher than 90% and 27 institutions had F1-scores higher than 80%. Table 1 shows the average F1-score for the dominating group among models trained on KimiaNet and DenseNet deep features. These results show a 16% and 22%

increase in F1-scores in the dominating and less-dominating groups compared to their counterparts trained with DenseNet's deep features. This finding substantiates the hypothesis that KimiaNet may have inadvertently learned to distinguish tissue source sites that could negatively affect the learning capability of an algorithm. Moreover, these data also suggest that it is plausible to distinguish source sites without an elaborate training process.

Table 1

Comparison of average source site classification performance between KimiaNet's and DenseNet's deep features. It suggests that KimiaNet's deep features contain information pertinent to the tissue source institution.

Tissue Source Site Institution	f1-score KimiaNet	f1-score DenseNet	Difference	Number of projects
Indivumed Inc	0.99	0.98	0.01	8
Memorial Sloan Kettering Cancer Center	0.93	0.82	0.11	18
Asterand Bioscience	0.91	0.83	0.08	17
University Health Network	0.91	0.62	0.29	7
Mayo Clinic	0.91	0.70	0.21	9
Barretos Cancer Hospital	0.89	0.61	0.28	7
ILSbio	0.88	0.82	0.06	15
MD Anderson	0.88	0.73	0.15	17
Case Western	0.86	0.72	0.14	2
Johns Hopkins	0.84	0.68	0.16	5
Erasmus MC	0.84	0.67	0.17	4
Roswell Park	0.84	0.68	0.16	18
University of California San Francisco	0.80	0.48	0.32	8
International Genomics Consortium	0.79	0.61	0.18	23
University of Pittsburgh	0.79	0.60	0.19	19
Cureline	0.78	0.56	0.22	10
University of North Carolina	0.75	0.48	0.27	24
Greater Poland Cancer Center	0.70	0.51	0.19	5
Duke University	0.69	0.52	0.17	5
Walter Reed	0.67	0.54	0.13	1
Henry Ford Hospital	0.64	0.32	0.32	2
Christiana Healthcare	0.64	0.50	0.14	17
University of Michigan	0.61	0.29	0.32	6
Emory University	0.41	0.24	0.17	5

An additional hypothesis that may explain the superior performance of KimiaNet versus DenseNet for tissue source institution detection is that KimiaNet can detect source sites because it distinguishes cancer subtypes, and not the other way around. However, it seems unlikely that the distribution of cancer subtypes among tissue source sites is not overly biased. Nevertheless, we tried to train and test for tissue source sites within each of the 30 TCGA projects/cancer types, for detailed results please see **Supplementary Table 2**. On average, models trained on KimiaNet's features were 14% more accurate than their counterparts trained on DenseNet's features. The former set of models were more accurate than the latter within every cancer type, ranging from 1–25%. Classification accuracies for 30 repeats within 5 cancer types with the highest number of WSIs are shown in Figure 2.

Bias and Image Search

In order to investigate the impact of institution-oriented patterns potentially discovered during feature extraction of digital pathology slides, the image search results are presented in this section. Although the deep networks describe extracted features from histopathology images based on a loss function defined on cancer subtypes, the information apparently is biased toward each contributing site of these digital slides. In this series of experiments, we match each query WSI with all WSIs with the same type of primary diagnosis to find the k most similar images. According to the number of hospitals and the number of samples from each institute, an Expected Value (EV) of k can be calculated to show how much we expect that k most similar images will be selected from the same site to which the query image belongs. For this purpose, each sample from the site H_i is searched among all samples to find the k most similar images (i.e., through minimum distance calculations). Then among the search results, we count the number of samples which are selected from H_i . Finally, this number is compared to EV to discover the possible internal bias toward the institution. In addition, the EV of the search results for each institution is calculated for all samples of each institute.

Figure 3 shows the search results on WSIs of six tumor types for three institutions with a maximum number of samples. For each institution, the bar plot indicates the number of WSIs found from the corresponding institution while searching for similar images. For instance, on Lung Squamous Carcinoma (LUSC) with 43 WSIs, there are 8 samples from Asterand Bioscience (Detroit, Michigan, USA). The five most similar images from those WSIs are determined among all LUSC samples. Therefore, $8 \times 5 = 40$ images will be the number of images in overall search results. According to the presented plots, 32 out of 40 images are selected from Asterand Bioscience, which is much higher than the corresponding EV (i.e., 5.81). This suggests the existence of bias towards the patterns extracted based on institution distinction rather than histomorphology of tumor types. In other words, the deep network has discovered hidden patterns in images that distinguish institutions. Consequently, the most similar images of a query image will predominantly be the WSIs from the same site the query WSI originates from.

Discussion

Participating medical institutions did not contribute equal numbers of cases to TCGA projects (cancer types). Additionally, samples provided by each institution have similar clinical characteristics (e.g., patient demographics, pathology diagnosis). Furthermore, WSIs provided by some institutions may have unique identifiers that could reveal which institution submitted the data. Convolutional neural networks may inadvertently pick up these identifiers or patterns even without any explicit fine tuning. For instance, DenseNet's features almost perfectly detected WSIs that originated from Princess Margaret Hospital and Indivumed Inc yielding F1-scores of 0.94 and 0.98, respectively. Tissue staining techniques may partly explain this phenomenon as WSIs provided by Indivumed Inc have similar color spectrums regardless of their organ of origin and cancer type (Figure 4). Such discrepancies could prompt deep models trained on TCGA data to learn tissue source site identifying patterns instead of, or at least in addition to, histopathological morphologic patterns. This was observed when KimiaNet picked up information that is useful for tissue source detection. For instance, we observed a drastic increase in detection performance for institutions such as Barretos Cancer Hospital, University Health Network, Mayo Clinic, and Moffitt Cancer Center. This finding implies that KimiaNet may have learned to distinguish source site institutions as a form of undesired (biased) shortcut to classify cancer types. It seems there are complex properties that could be attributed to WSIs provided by each institution, such as slide preparation protocols, hematoxylin and eosin (H&E) stain variation, proprietary image details, and characteristic image noise (see examples in Figure 4). Additionally, convolutional layers may easily collect irrelevant noisy patterns of input images, see Figure 4c. Based on our results, it seems that such patterns could reach deep into the network up until the last layers that we use for image representation.

Although KimiaNet seems to have learned some non-morphologic tissue source site patterns, this is dissimilar to the classic “wolf-versus-husky” problem,²² where a deep network learns to distinguish wolves by just detecting the snow in the background of the image. KimiaNet still needs to learn histologic patterns to distinguish various cancer types provided by a certain institution. For example, it seems KimiaNet can accurately identify WSIs of Memorial Sloan Kettering Cancer Center, which contributed samples for 18 various cancer types. Therefore, the knowledge about the source institution does not reveal the diagnosis but helps to exclude some cancer types. If we assume that such loops exist in the decision-making process of deep networks, they may cause problems in their deployment for clinical utility. If a model mistakenly or correctly attributes a sample to a certain hospital, the reliability of the network's output will drop. This is not something easily preventable, as convolutional layers easily pick up image noise due to the usage of large number of filters they use. As well, staining techniques may be ineffective in medical institution obfuscation.

The most apparent problem when dealing with deep networks in medical image analysis is an apparent overestimation of their performance when trained on TCGA WSIs and perhaps other large WSI repositories. This issue usually manifests itself in a significant performance drop revealed by external validation. Hence, it is advisable to divide training and test data based on the tissue source sites when using TCGA data in order to prevent the model from seeing samples from some institutions and therefore creating a virtual external validation set within TCGA. Although such suggestions may help researchers

better estimate their model's performance without going through the trouble of finding external validation data, it does not address the main issue: deep models may find detours toward easily detectable irrelevant patterns to achieve their goals. Even after excluding some medical institutions from the training data, model learning was impacted by existing biases. It has also been shown that it is not easy to eliminate tissue source site biases through stain normalization techniques.²⁰ Therefore, the real challenge is to find ways to prevent models from learning transiently useful but histologically irrelevant information. It has also been shown that the distributions of some clinical parameters such as tumor stage, gene expression, receptor status, and survival prognosis are significantly different across tissue source sites. Our finding that a deep network learned to accurately distinguish tissue source sites even though it had not been trained to do so raises the possibility that some research papers published about predicting clinical information from histopathology images without evaluation on external validation data may have been using irrelevant features. One can argue that it is not enough to evaluate deep models on external validation data; we should devise deceptive tests to estimate the robustness of deep models. To give an analogy, it is useful for a handwritten digit recognition model to gain knowledge of who wrote the letter, as everybody's handwriting has unique subtleties. By exploiting this information, the model would perhaps oversimplify the problem. In extreme cases, this supposed "cheating" prevents the model from recognizing the intrinsically relevant and sensible patterns. Perhaps what such models are learning is tantamount to creating a flawed lookup table with limited (i.e., non-generalizable) success.

Bias in healthcare emerges when an AI algorithm takes into account existing social inequities in such as race, ethnicity, socioeconomic in its decision-making process. Therefore, such biased models could eventually exacerbate inequities in healthcare, if being widely implemented.²³ There are examples of AI algorithms being biased toward race or gender in non-healthcare situations,^{24–27} such as recidivism prediction algorithm that was significantly biased against African American defendants.²⁸ Since hospitals are located in areas with various demographic characteristics, identification of hospitals not only could cripple a model's generalizability, but also could introduce social biases into the model's decision-making process, which would consequently result in the models becoming biased towards demographic characteristics with far-reaching detrimental social consequences.

In summary, this study showed that TCGA images have unique identifiers that could be used to recognize their respective tissue source hospitals. Convolutional neural networks seem to pick up these identifiers without the need to be explicitly trained to do so. Additionally, a CNN trained for cancer subtype classification has mastered tissue source site identification during its learning phase, and perhaps uses this additional (but irrelevant) information for cancer subtype classification. This phenomenon can negatively affect WSI representation for image classification and search, where categorized images were seen to be disproportionately coming from the same hospital to that of the query image.

Methods

Determining the Tissue source medical institution. All TCGA samples have a unique code starting with "TCGA", followed by a two-character code called tissue source site (TSS), which is unique to the

institution and cancer type. For example, TSS code 02 is for Glioblastoma (GBM) samples provided by MD Anderson. Descriptions of TSS codes are available at this URL address, <https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes>, last accessed on 13/08/2021.

For all diagnostic paraffin-embedded hematoxylin and eosin (H&E) slides, TSS codes were obtained to determine the source institutions. In some cases, naming was inconsistent, therefore we replaced some names. For instance, Memorial Sloan Kettering sometimes was called MSKCC; we replaced MSKCC instances with Memorial Sloan Kettering. We acknowledge that there might still be some inconsistency in institutional naming despite our efforts.

Tissue patch extraction process. As per Yottixel search paradigm,¹⁵ WSIs were clustered into 9 clusters at 5x magnification based on RGB histograms. As well, tissue patches of size 1000x1000 pixels at 20x magnification were selected proportionally to the size of each cluster. Tissue patches with low cellularity were discarded to increase the ratio of tissue patches extracted from malignant regions. More details can be found in the paper published by Riasatian et al.¹⁹

Deep feature extraction. The outputs of the last pooling layer of KimiaNet and DenseNet121 were separately used for tissue patch representation. These two networks have the same structures, but the former was trained for cancer type classification on TCGA data. Tissue patches were fed to these two networks to extract deep feature vectors of size 1024 for each patch.

Tissue source institution determination. Deep feature vectors were subsequently fed to a small network with two fully connected hidden layers of size 500 and 200 neurons with a rectified linear unit (ReLU) activation function. The models were trained for 5 epochs with a batch size 60, Adam optimizer, and sparse categorical cross-entropy as the loss function.

Image Search. To undertake the search, we need to calculate the distance between pairs of images represented through their deep features. However, since each WSI is made of several tissue patches, a patch-wise method is required to calculate the distance between the feature vectors of a WSI; a “median-of-min” approach is conducted for this purpose.¹⁵ To calculate the distance between a query image and another WSI, the minimum distance between each patch of the query image and all patches of other WSI is computed first. The median value of all minimum distances was taken as the final distance between the query and compared WSIs. Thus, the most similar images with minimum distance are accordingly identified.

Declarations

Author's contributions

T.D., A.A. and P.R.carried out the experiments. A.R. provided the data.H.T. and S.R supervised the project. C.C. and L.P. verified the results and the interpretations. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

References

1. Gutman, D. A. *et al.* Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J. Am. Med. Informatics Assoc.***20**, 1091–1098 (2013).
2. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkol.***1A**, 68–77 (2015).
3. Liao, H. *et al.* Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clin. Transl. Med.***10**, (2020).
4. Tabibu, S., Vinod, P. K. & Jawahar, C. V. Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Rep.***9**, 10509 (2019).
5. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.***24**, 1559–1567 (2018).
6. Iizuka, O. *et al.* Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci. Rep.***10**, 1504 (2020).
7. Aatresh, A. A., Alabhya, K., Lal, S., Kini, J. & Saxena, P. P. LiverNet: efficient and robust deep learning model for automatic diagnosis of sub-types of liver hepatocellular carcinoma cancer from H&E stained liver histopathology images. *Int. J. Comput. Assist. Radiol. Surg.* (2021) doi:10.1007/s11548-021-02410-4.
8. Chen, M. *et al.* Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *npj Precis. Oncol.***4**, 14 (2020).
9. Martino, F. *et al.* Deep Learning-Based Pixel-Wise Lesion Segmentation on Oral Squamous Cell Carcinoma Images. *Appl. Sci.***10**, 8285 (2020).
10. de Bel, T. *et al.* Automatic segmentation of histopathological slides of renal tissue using deep learning. in *Medical Imaging 2018: Digital Pathology* (eds. Gurcan, M. N. & Tomaszewski, J. E.) 37 (SPIE, 2018). doi:10.1117/12.2293717.
11. Jang, H.-J., Lee, A., Kang, J., Song, I. H. & Lee, S. H. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J. Gastroenterol.***26**, 6207–6223 (2020).
12. Wulczyn, E. *et al.* Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One***15**, e0233678 (2020).
13. Wan, J. *et al.* Deep Learning for Content-Based Image Retrieval. in *Proceedings of the 22nd ACM international conference on Multimedia* 157–166 (ACM, 2014). doi:10.1145/2647868.2654948.
14. Komura, D. *et al.* Luigi: Large-scale histopathological image retrieval system using deep texture representations. *bioRxiv* (2018) doi:10.1101/345785.

15. Kalra, S. *et al.* Yottixel – An Image Search Engine for Large Archives of Histopathology Whole Slide Images. *Med. Image Anal.***65**, 101757 (2020).
16. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009). doi:10.1109/CVPR.2009.5206848.
17. Schaer, R., Otálora, S., Jimenez-del-Toro, O., Atzori, M. & Müller, H. Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. *J. Pathol. Inform.***10**, 19 (2019).
18. Schaumberg, A. J. *et al.* Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Mod. Pathol.***33**, 2169–2185 (2020).
19. Riasatian, A. *et al.* Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Med. Image Anal.***70**, 102032 (2021).
20. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **2021 12****112**, 1–13 (2021).
21. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.***3**, 610–619 (2021).
22. Ribeiro, M. T., Singh, S. & Guestrin, C. ‘Why Should I Trust You?’ in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016). doi:10.1145/2939672.2939778.
23. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health***9**, 10318 (2019).
24. Tian, J., Xie, H., Hu, S. & Liu, J. Multidimensional Face Representation in a Deep Convolutional Neural Network Reveals the Mechanism Underlying AI Racism. *Front. Comput. Neurosci.***0**, 17 (2021).
25. Ntoutsis, E. *et al.* Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Min. Knowl. Discov.***10**, e1356 (2020).
26. Leavy, S. Gender bias in artificial intelligence. in *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* 14–16 (ACM, 2018). doi:10.1145/3195570.3195580.
27. Zou, J. & Schiebinger, L. AI can be sexist and racist – it’s time to make it fair. *Nat.* **2021 559****7714****559**, 324–326 (2018).
28. Khademi, A. & Honavar, V. Algorithmic Bias in Recidivism Prediction: A Causal Perspective (Student Abstract). *Proc. AAAI Conf. Artif. Intell.***34**, 13839–13840 (2020).

Figures

Figure 1

The overview of the training process. WSIs were divided into test and training sets. In our study, 619 patients had more than one WSIs, which were excluded from the test set and only used for training. Tissue patch samples of size 1000 x 1000 pixels derived from WSIs were fed into deep networks (KimiaNet and DenseNet) for feature extraction. The output of the last pooling layer was used as deep features. The extracted deep features were later used to train a smaller network with two hidden layers to identify tissue sources.

Figure 2

The boxplot of source site accuracies of DenseNet's and KimiaNet's deep features over 30 repeats within 5 cancer types with the highest numbers of WSIs. One can see that models trained on KimiaNet's deep features are consistently more accurate than their counterparts in DenseNet. This finding suggests that KimiaNet's deep features contain information about source sites of WSIs, although it was originally trained to distinguish cancer types and not source sites. It seems that this additional information, perhaps medically irrelevant, helps the network to classify cancer types due to the TCGA dataset's internal biases.

Figure 3

The plots of search results to verify the existence of bias in data. The green bar plots represent the overall number of images in search results from the institute which the query image belongs to. Whereas the orange plots indicate the expected value of this quantity. The number of samples from each institute is presented in the label of the institute. In most of the institutes, the calculated number is more than 4 times higher than the EV.

Figure 4

Potential explanations of how deep networks identified tissue source institutions even without explicit training. a) Tissue patches of WSIs provided by the institution “Indivumed Inc” were accurately identified by using DenseNet’s and KimiaNet’s deep features, perhaps due to their unique stain intensities and color spectrum. One can observe that they have a dominant purplish stain. b) Digital pathology images are generally noisy. The noise could be amplified by histogram equalization techniques. Although those noisy patterns are more prominent in the background, they most likely exist over the tissue area as well. c)

Convolutional layers could easily pick up noisy patterns existing in a background patch, and over-emphasizing them, as shown by visualizing the output of a convolutional layer of KimiaNet.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)