CHAPTER

3

Deep learning for biomedical videos: perspective and recommendations

David Ouyang, Zhenqin Wu, Bryan He and James Zou

Abstract

Medical videos capture dynamic information of motion, velocity, and perturbation, which can assist in the diagnosis and understanding of disease. Common examples of medical videos include cardiac ultrasound to assess cardiac motion, endoscopies to screen for gastrointestinal cancers, natural videos to track human behaviors in population health, and microscopy to understand cellular interactions. Deep learning for medical video analysis is rapidly progressing and holds tremendous potential to extract actionable insights from these rich complex data. Here we provide an overview of deep learning approaches to perform segmentation, object tracking, and motion analysis from medical videos. Using cardiac ultrasound and cellular microscopy as case studies, we highlight the unique challenges of working with videos compared to the more standard models used on still images. We further discuss available video datasets that may search as good training sets and benchmarks. We conclude by discussing the future directions for this field with recommendations to practitioners.

Keywords: Video; deep learning; echocardiogram; microscopy; segmentation; motion analysis

3.1 Introduction

Artificial intelligence and machine learning has seen dramatic advances in the last 10 years. While the concepts of neural networks, convolution operations, and methods to train networks have been proposed over the last 30 years,[1] it is relatively recent the widespread availability of graphic processing units and the insight that this hardware can efficiently perform the repetitive, parallel operations that speeds up the training of these complex machine learning algorithms.[2] With these advances in computation, deep neural networks, in which many layers of mathematical operations are used to learn complex relationships, have been used to tackle complex tasks including genomics,[3] computer vision,[4] natural language processing,[5] and human strategy games.[5,6] This field of "deep learning" has seen some of the most exciting accomplishments in artificial intelligence.

DOI: https://doi.org/10.1016/B978-0-12-821259-2.00003-X

Many of the biggest advances and best examples of the high performance of deep learning have been in computer vision, the scientific field of designing computer systems to understand images and videos.[2,5,6] While computer vision has been showing steady, incremental improvement over many years, the introduction of deep convolutional neural networks to the task of image classification produced drastic improvements and highlighted the potential of deep learning compared to previous state-of-the-art techniques using feature engineering.[6,7] Paralleling the progress with still images, deep learning on video datasets have been shown to have outstanding results by incorporating elements of neural network architectures originally tailored for both still image computer vision and time series data.[8−11]

Inspired by the tremendous, near human level of accuracy in classifying images, researchers attempted to apply similar deep learning algorithms to medical imaging tasks.[12,13] Ranging from pictures of skin lesions in photographs and retina images from ophthalmologists to chest X-rays and mammograms, deep learning approaches have been adapted to medical still image datasets.[12−17] While these are complex datasets, an even richer set of data exists in medical videos, which capture motion and behaviors that still images cannot detect. In one particularly salient example, the cardiovascular system has many dynamic structures, with the motion of heart muscle, heart valves, and blood providing significant diagnostic information that still images do not capture.

Deep learning for medical videos is much less advanced compared to deep learning for images, though it holds tremendous potential for impact. In this chapter, we review key advances in computer vision and deep learning on video tasks and highlight applications in the medical machine learning literature. We discuss case studies ranging from cardiac ultrasounds (echocardiograms) to microscopy videos, which highlight approaches to understand dynamic systems and techniques to tackle complex datasets.

3.2 Video datasets

A key factor in the advances of computer vision has been the machine learning community's adoption of standardized datasets and comparison benchmarks for the evaluation of machine learning algorithms (Fig. 3.1). These datasets, which often comprise the imaging data with human annotations, dictate the scope of tasks that are being answered. In the still image dataset realm, ImageNet is a high profile example of a large imaging dataset used to benchmark and study the relative performance of machine learning models.[7] A large dataset of natural images obtained from Google Image Search with crowdsourced classifications, ImageNet was a standardized input dataset that could be used in competitions. The drastically improved performance of deep learning algorithms on ImageNet classification was one of the first hints of the potential of convolutional neural networks.[2,7]

Video data are frequently used in medicine and biological sciences. Natural human videos are used to study behaviors, record interviews, and track actions. Research in mental health and neuroscience often relies on the recording of human behavior in video format. In the hospital the diagnosis of epilepsy incorporates patient videos as well as electrical recordings of brain activity. The diagnosis of many diseases requires the observation of behavior and motion (or limitations in motion), and many physical exam

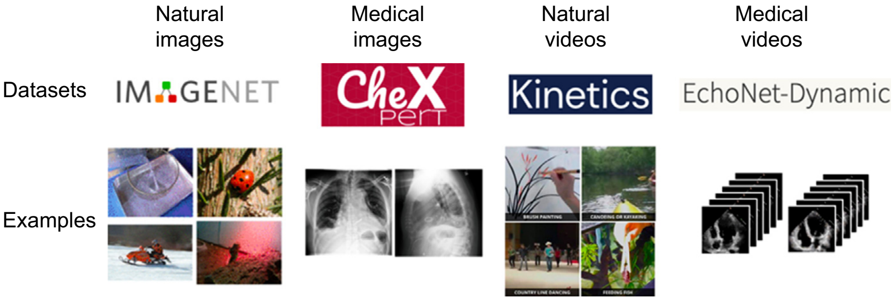

FIGURE 3.1 Examples of publicly available datasets with representative frames.

maneuvers performed by physicians seek to elicit differences in behavior with perturbation or stress. Unfortunately, many of these medical visual behaviors are not consistently recorded, which makes machine learning for these tasks more difficult. As with all tasks within machine learning for health, understanding the clinical scenario and available datasets inform the machine learning pipeline and possible training tasks.

Videos have additional temporal information compared to still images, and many tasks require this information for understanding and comprehension. While any individual frame of a video can give information on the location and context, many behaviors and movements require comprehending temporal information. For example, while a still image might be enough to identify a door, the video is required to understand whether the action consists of "a door closing" or "a door opening." Biological behavior is very complex, often consisting of a similar actor but different motions or actions which dictate the task at hand. Actions such as "patting a person's head" versus "braiding hair" can appear visually similar in a still image, but the temporal information encoded in a video data can readily distinguish between different behaviors. Datasets such as Kinetics,[18] HMDB,[19] and UCF101[19,20] have been designed for the purpose of investigating computer vision on human behavior videos.

There are also many forms of advanced medical imaging that capture motion for disease diagnosis. An example of understanding motion for medical diagnosis is the imaging of the heart for detecting cardiovascular disease. The heart is a tremendously dynamic organ, with motion in every heartbeat and often sizable variation even beat-to-beat. While the heart can be imaged through many modalities—including ultrasound (echocardiograms), computed tomography (CT), or magnetic resonance imaging (MRI)—modalities which have lower temporal resolution often require aggregation of information and taking advantage of the cyclical nature of the cardiac cycle and each heartbeat. Thus abnormalities in the heart muscle or aberrations in heart valve function can be readily detected in multiple imaging modalities, but all modalities of cardiac imaging incorporate the temporal information.

EchoNet-Dynamic is an example of a publicly available medical video dataset.[14] Comprised of over 10,000 echocardiogram, or cardiac ultrasound, videos and associated expert labels of heart chamber sizes and cardiac function, the EchoNet-Dynamic dataset was released to the machine learning for health community to serve as a benchmark for medical video research and evaluation of domain specific architectures. By having a standardized shared dataset, direct "apples-to-apples" comparison of different machine learning models can be performed on key medical questions. Both clinically relevant and taking advantage of the information specifically encoded in video, cardiac function is a challenging benchmark clinical task that will advance machine learning in healthcare.

With all medical datasets, concerns regarding patient privacy, fairness, and generalizability need to be addressed. Unlike synthetic data or natural images, medical data can come from especially vulnerable populations, be biased toward certain demographics which can cause bias to be propagated in machine learning model trained on the dataset, and might not be generalized to the population as a whole.[21] Often, special efforts need to be made prior to data release to verify the balance of the dataset, avoid bias in patient selection, and removal of identifying features or markers whenever possible.[22] In the case of EchoNet-Dynamic, in addition to evaluating the demographic information of the patient population, each video was manually reviewed by a trained employee of the hospital system to highlight and exclude identifying information.

With video datasets, many different important clinical tasks can be performed. In the next few sections, we will use examples from healthcare and biological research to showcase cases and models used for semantic segmentation, object tracking, and motion classification.

3.3 Semantic segmentation

Semantic segmentation refers to the task of labeling each individual pixel of an image or video with corresponding labels, often of classification tasks to identify regions and structures (Fig. 3.2). In the natural world, humans perform this task instinctively, seamlessly identifying objects such as cars, people, or bicycles in order to interact with the natural environment. Prior to taking an action (e.g., getting in the car), one needs to recognize where the car is and how it is oriented. This task is crucial for understanding the local environment and is more difficult than traditional object-identification tasks, which often simply ask if an image contains a certain object but not necessarily ask where the object is in the image.

The machine learning community has released datasets such as Microsoft Common Objects in Context (COCO)[14,23] and Cityscape Dataset for Semantic Urban Scene Understanding[24] for studying the machine understanding of local environments and the pixel-wise relationship between regions and objects. Outside of health-care applications the potential of self-driving cars and other disruptive technologies motivate research into the understanding of urban traffic environments through semantic segmentation of videos.

In medicine, characterizing organ systems through medical imaging often relies on similar approaches to understanding the voxel-wise relationship between medical imaging and disease characteristics. For example, in the example of solid organ cancers, such as prostate cancer and lung cancer, radiologists take significant time to understand the size and distribution of tumors. The tumors are often distinct and readily recognizable; however, the clinical workflow has much subjectivity and human variation in how to measure

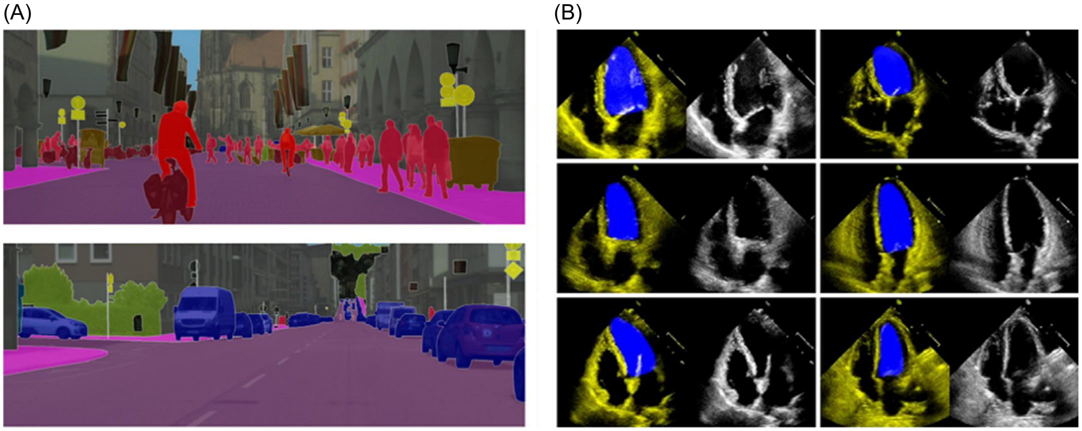

FIGURE 3.2 Semantic segmentation task examples in natural video and medical video. (A) The Cityscape Dataset for Semantic Urban Scene Understanding identifies common physical structures and classes in an urban commuter environment. (B) The EchoNet-Dynamic Dataset identifies left ventricular size and shape to characterize cardiac function.

the dimensions and characteristics of the tumor. Such tasks are crucial in distinguishing between dormant versus progressive disease; however, human variation can lead to the under diagnosis of subtle changes in tumor burden and neglect small but meaningful changes in disease state.

Many prominent and well-studied neural network architectures have been designed with semantic segmentation tasks in mind. Fully convolutional networks (FCNs),[25,26] U-Net,[27] and DeepLab[28,29] represent the gamut of architectures and designs evaluated for semantic segmentation tasks in computer vision datasets. Common to all listed architectures is a model design that aggregates both distant and local information from other pixels and collapses pixel-wise information into a smaller vector or array that more closely represent the labeled tasks and reexpanding that annotation into an array of the original size that represent the pixel-level label. An example of an application driven architecture which has subsequently been expanded to other biomedical as well as nonmedical tasks is the U-Net architecture, which was originally inspired by the task of segmenting electron microscopy images to annotate cells and small local structures. This architecture extends advances made with FCNs by aggregating input image information into smaller and smaller layers that encode higher order information (which has been described as the "encoding arm") followed by gradual upscaling of layer sizes to produce an output of the same shape and size as the original input image (described as the "decoding arm"). Given its efficiency and high performance, U-Net has been applied to many tasks outside of biomedical imaging and has been recognized as an advance for machine learning computer vision even outside of medicine.[30]

Many semantic segmentation models rely heavily on individual frame level information while discounting additional information for temporally adjacent frames. In nonmedical datasets such as CityScapes the dataset is often constructed with sparse labeled frames selected to maximize differences between sampled frames rather than providing labels for multiple consecutive frames to augment the model training. In the example of echocardiogram videos, researchers were able to show that training on a small subset of frames was able to generalize the entirety of the video and the previously unannotated frames.[14] In this example, opportunities in augmenting the dataset exist knowing the constraints of the particular data. For example, for echocardiograms, sonographers often trace the left ventricle at its largest and smallest, so the shape and size of the ventricle is smoothly constrained between the two examples even in unlabeled frames. This allows fuzzy model training by using the same training labels to train adjacent frames of the video as well as penalizing the model when there is drastic change in model prediction from frame to frame. Techniques from still images can be generalized to video and from a nonmedical domain to a medical domain.

3.4 Object detection and tracking

A common subsequent task after image segmentation is the detection, classification, and tracking of objects. Understanding the mechanisms of basic biological processes requires understanding the prevalence, location, and trajectory of cells and subcellular objects in molecular and cellular imaging. For example, our understanding of chemotaxis in immunology comes from studying the gradual movement of neutrophils and experimental modifications that slow or impede cellular movement. Given the diversity and high numbers of cellular actors in complex biological processes, automated approaches to detect, classify, and track important objects in cellular videos is crucial to advance our understanding fundamental biological processes. Deep learning has revolutionized how we can computationally detect and track objects and have made leaps in how we interpret cellular videos.

In addition to differentiating between foreground and background, computer vision tasks involving visual information requires the understanding of the objects represented in the image and video. In cellular microscopy, similar cell types and populations can be represented by many morphologies, shapes, and sizes. Machine learning models must understand intrinsic characteristics that define different cellular populations to detect and track cellular movement. Understanding of the various projections an individual object can manifest in visual information from different views and perspectives is important in appropriately detecting and identifying objects in visual data. Often this is made even more computationally difficult in biomedical imaging, in which a single field of view can have many, even hundreds of instances of the objects of interest. Appropriate object registration and detection is needed to analyze the population level motion and trajectories.

In the recent development from deep learning community, the task of object detection/instance segmentation is solved through a complex framework which involved multiple prediction stages and heads that are responsible for generating proposals, and identifying bounding boxes and labels of objects. Some well-known models include RCNN,[31] Mask-RCNN,[32] and YOLO.[33] These models are typically trained in a supervised learning framework with matching images and labels in the form of annotated bounding boxes. In biomedical data, especially cell imaging, usage of these advanced neural network architectures is typically limited due to lack of human annotations, heterogeneity in human labels, and the large amount of relatively homogeneous objects in images that make human annotation tedious.

Alternative solutions have been proposed which generalize from the pixel map output from semantic segmentation and perform heuristic-based instance separation. Each segmentation can be classified into relevant cell types, although an important biological challenge from cellular imaging is the challenge of segmenting large clusters of often overlapping objects. In cases when cells are distinct from each other, the semantic segmentations can be sufficient. However, when there are overlapping cell pairs or groups of closely contacting cells, even little segmentation error could lead to classification errors and fused cell instances. A large variety of segmentation methods have been proposed to answer the challenge of separating close, overlapping objects in biomedical imaging. Some existing solutions rely on an assumption of cell shape, employing strategies such as Laplacian of Gaussian,[34] radial symmetry transformation,[35] and distances between objects to distinguish between individual instances. Other researchers have tried feature-based or threshold-based methods to distinguish between different objects.[36] The wide range of proposed techniques highlight the challenge of cellular object detection and each method has shown impressive results in specific imaging modalities and domain questions (Fig. 3.3).

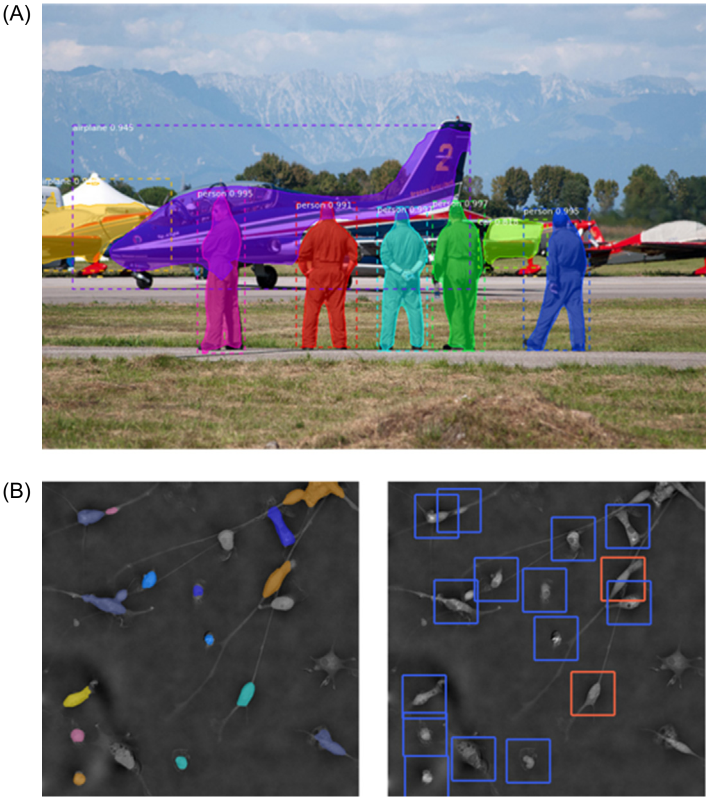

FIGURE 3.3 Object detection and instance segmentation task examples (A). Mask-RCNN generates bounding boxes and masks for objects in natural images (B). Cell detection in microscopy images.

Video data provides additional temporal information that can be used to inform object detection and track objects even through challenging frames that independently might be difficult to process. The additional information can also make object tracking more difficult—particularly when there are multiple objects in the same frame and pairing each object with the same object in a preceding or subsequent frame is needed to accurately generate trajectories. This matching task is usually solved under the framework of linear assignment problem[37]: matching a set number of objects from one frame of a video to the same objects in a subsequent frame. In this problem setting a cost matrix is specified based on how likely a pair of cells in two image frames are from the very same cell. This matrix is usually defined/calculated based on factors including the distance of locations, similarity of appearance, and surrounding environment. Though in most cases these components are empirically selected and weighed in the final cost matrix, there is recent work applying neural network-based method to refine the cost matrix composition.[38,39] Further refinement can be performed after generating an initial set of trajectories from frame-to-frame matchings.[37] This step can mitigate segmentation error and also account for events such as cell merging and splitting.

In addition, deep learning can help researchers investigate the complex relationship between visual phenotype and genetics. Convolutional neural networks[28] provide a powerful and unbiased tool to organize and quantify the complex morphological characteristics of cells from imaging data. Morphological and morphodynamic states are often highly correlated with gene expression—it could be easily imagined that a convolutional network-based featurizer will be able to extract the relationship between the visual phenotype and the gene expression and functional states of cells. Cellular videos enable a wide range of studies on time-dependent behavior in biological systems that are not possible with traditional microscopy. There is increasing recognition that cellular systems are incredibly dynamic[40] throughout the cell cycle[41] and further information can be obtained in dynamic analysis of the cell morphology.[42] In a live cell imaging, tracking trajectories of cells opens opportunities for detailed analysis on dynamic state and temporal change of individual cells during development and in immunological processes.

3.5 Motion classification

In both natural video as well as in medical imaging machine learning, there are tasks that require understanding motion and the interplay of structures. Image-based classifiers traditionally have a tough time distinguishing between opening or closing a door, or repetitive motions such as brushing or braiding hair. In medicine a large range of tasks require understanding of biological movement. For natural videos, subtle physical motions and variation can identify important physiological states and medical diagnoses.[43–45] In healthcare, video medical imaging is obtained for the diagnosis of cardiovascular disease particularly because the heart is a dynamic structure and abnormalities in the heart muscle and valves is most clearly reflected in abnormal heart motion. In this section, we present examples of machine learning applied to assessing cardiac function to exemplify opportunities with video-based deep learning architecture.[14]

The "convolution" part of convolutional neural networks refers to the collections of mathematical functions that aggregate and pass information into subsequent layers of the neural network.[2,7,14] For image-based computer vision tasks the convolutional task aggregates local pixel information and benefits from embedding understanding of local structure and geometric shape.[29] In still image tasks, all visual information is mapped to a two-dimensional (2D) data structure and the relevant functions are 2D convolutions that traverse through the image for an output often of similar size and structure. While there has been sizable advances in neural network architecture design,[9,28,29,46] the fundamental mathematical operation is 2D convolutions for processing still images.

Videos contain temporal information, and this richer dataset can be represented in many different ways, and the neural network architectures used to understand them are similarly more complex.[8,10,18] Video data can be preprocessed with feature-tracing algorithms such as optical flow and dense optical flow to consolidate information and label regions with motion and activity prior to classification.[47] Other researchers have treated the temporal information as another dimension in the data, such that a three-dimensional (3D) data structure (x,y,z) represents the input video with x and y axes representing the spatial information of each frame of the video and the z axis representing temporal information across frames.[10,47] In this approach, 3D convolutional kernels are used to consolidate information from all three dimensions, and both spatial and temporal information is integrated in the model predictions. In order to minimize computational cost, various approaches to independently treat temporal and spatial information have been attempted with good efficacy.[10]

In cardiac physiology the vigor and speed of contractions of the heart chambers, especially the left ventricle, is quantified to understand cardiac function. Human physicians use video information from cardiovascular ultrasound (echocardiograms), CT, or MRI to examine the heart. Severe impairment of heart muscle movement captured on these medical imaging videos is considered "heart failure" and is the leading cause of hospitalization in the United States.[48] Researchers have applied video-based deep learning models to predict heart function with high accuracy and precision.[10,14] Even human interpretations of heart motion can be subjective and vary among experts,[49,50] and appropriately applied deep learning models can improve the reliability in medical imaging and physician and patient trust in diagnostic testing.

3.6 Future directions and conclusion

There have been significant advances in machine learning applied to medical imaging and medical videos. From applying conventional computer vision algorithms to standard medical imaging to proposing novel deep architectures inspired by biomedical imaging segmentation tasks, medical imaging applications has inspired basic machine learning research and machine learning is on the verge of revolutionizing how medical imaging is interpreted. Many further prospective studies need to be performed, and the interplay between human elements and machine learning models need to be understood before deep learning can be truly applied in a clinical setting, but the future is full of opportunities for machine learning in healthcare.

The various ways humans interact with natural image and video data directly correspond to opportunities to standardize, improve, and expand availability to medical imaging through machine learning. While not explicitly defined in these categories, image segmentation, object detection, object tracking, and motion analysis are used daily in medical imaging by human radiologists, cardiologists, and pathologists to understand and diagnose disease. In a combination of these discrete tasks, cardiologists identify different chambers of the heart and track heart motion to assess cardiac function. Video-based features, such as desynchrony or impairment of motion, defines cardiovascular disease such as electrical conduction abnormalities and cardiomyopathy.

With the gravity of medical decision-making, it is likely the first application of video artificial intelligence (AI) models will be on replacing tedious, unglamorous intermediate tasks rather than providing end-to-end predictions of a final medical diagnosis. In the example of echocardiograms the first application of video AI models might be using AI to label the left ventricle rather than directly replacing the human input and producing a final diagnosis. In many ways, this would already improve upon the current clinical workflow by using multiple cardiac beats to inform human diagnosis of cardiac function, but this would also parallel the current clinical workflow of initial tracing by sonographers or trainees before being signed off by a physician. However, additional work needs to be done to study what happens when AI models get good enough, but still need an observant overseer to maintain quality. For example, in the example of self-driving cars, one can foresee a future where AI systems can handle the vast majority of mundane experiences but still need human intervention in difficult environments—how does such a system keep a human engaged and situationally aware while maintaining human trust in the AI system? In the future, there needs to be further work to understand and assess the relationship between human and machine learning models when applying video AI models.

Important work is currently being done to have open, standardized, and shared datasets for studying and benchmarking machine learning of medical imaging. Implicit in this challenge is the desire to understand cross-institution differences in image acquisition and interpretation while also maintaining patient privacy and understanding implicit biases in the health-care system. Fairness in machine learning has become an increasingly important issue as we increasingly recognize that both implicit and explicit biases in the training labels and sampling of examples can significantly influence machine learning model behaviors. In medicine, biases in how physicians diagnose patients with different socioeconomic backgrounds and biases in the availability of health-care tests and resources can be propagated into machine learning models if careful examination and reflection is not undertaken.

References

1. LeCun Y, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1989;**1**:541−51.
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in neural information processing systems 25*. Curran Associates, Inc.; 2012. p. 1097−105.
3. Zou J, et al. A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12−18.
4. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278−324.

5. Devlin J, Chang M-W, Lee K, Toutanova K. *BERT: pre-training of deep bidirectional transformers for language understanding*. *arXiv [cs.CL]*; 2018.
6. Silver D, et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;**529**:484–9.
7. Russakovsky O, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015;**115**:211–52.
8. Tran A, Cheong L-F. Two-stream flow-guided convolutional attention networks for action recognition. *2017 IEEE international conference on computer vision workshops (ICCVW)* 2017. Available from: https://doi.org/10.1109/iccvw.2017.368.
9. Song L, Weng L, Wang L, Min X, Pan C. Two-stream designed 2D/3D residual networks with LSTMs for action recognition in videos. *2018 25th IEEE international conference on image processing (ICIP)* 2018. Available from: https://doi.org/10.1109/icip.2018.8451662.
10. Tran D, et al. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF conference on computer vision and pattern recognition* 2018. Available from: https://doi.org/10.1109/cvpr.2018.00675.
11. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: *Proceedings of the IEEE international conference on computer vision* 2015;4489–97.
12. Esteva A, et al. Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**546**:686.
13. McKinney SM, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;**577**:89–94.
14. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;**580**: 252–256. Available from: https://doi.org/10.1038/s41586-020-2145-8.
15. Bello GA, et al. *Deep learning cardiac motion analysis for human survival prediction*. *arXiv [cs.LG]*. 2018.
16. Poplin R, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;**2**:158–64.
17. Coudray N, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;**24**:1559–67.
18. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 2017;6299–308.
19. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion recognition. In: *2011 international conference on computer vision* 2011;2556–63.
20. Soomro K, Zamir AR, Shah M. *UCF101: a dataset of 101 human actions classes from videos in the wild*. *arXiv [cs.CV]*. 2012.
21. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature* 2018;**559**:324–6.
22. Kim MP, Ghorbani A, Zou J. Multiaccuracy. *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society – AIES '19* 2019. Available from: https://doi.org/10.1145/3306618.3314287.
23. Lin T-Y, et al. *Microsoft COCO: common objects in context*. *arXiv [cs.CV]*. 2014.
24. Cordts M, et al. The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016.
25. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *2015 IEEE conference on computer vision and pattern recognition (CVPR)* 2015. Available from: https://doi.org/10.1109/cvpr.2015.7298965.
26. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. *Semantic image segmentation with deep convolutional nets and fully connected CRFs*. *arXiv [cs.CV]*. 2014.
27. Dong H, Yang G, Liu F, Mo Y, Guo Y. *Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. Medical image understanding and analysis*. Springer International Publishing; 2017. p. 506–17.
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE conference on computer vision and pattern recognition (CVPR)* 2016. Available from: https://doi.org/10.1109/cvpr.2016.90.
29. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 2018;**40**:834–48.
30. Billaut V, de Rochemonteix M, Thibault M. *ColorUNet: a convolutional classification approach to colorization*. *arXiv [cs.CV]*. 2018.
31. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE conference on computer vision and pattern recognition* 2014. Available from: https://doi.org/10.1109/cvpr.2014.81.

32. He K, Gkioxari G, Dollar P, Girshick R. Mask RCNN. *2017 IEEE international conference on computer vision (ICCV)* 2017. Available from: https://doi.org/10.1109/iccv.2017.322.
33. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. *2016 IEEE conference on computer vision and pattern recognition (CVPR)* 2016. Available from: https://doi.org/10.1109/cvpr.2016.91.
34. Xu H, Lu C, Berendt R, Jha N, Mandal M. Automatic nuclei detection based on generalized Laplacian of Gaussian filters. *IEEE J. Biomed. Health Inf.* 2017;**21**:826−37.
35. Loy G, Zelinsky A. Fast radial symmetry for detecting points of interest. *IEEE Trans Pattern Anal Mach Intell* 2003;**25**:959−73.
36. Vicar T, et al. Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinform* 2019;**20**:360.
37. Jaqaman K, et al. Robust single-particle tracking in live-cell time-lapse sequences. *Nat Methods* 2008;**5**:695−702.
38. Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: learning to track multiple cues with long-term dependencies. *2017 IEEE international conference on computer vision (ICCV)* 2017. Available from: https://doi.org/10.1109/iccv.2017.41.
39. Moen E, et al. *Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning*. Available from: https://doi.org/10.1101/803205.
40. Kimmel JC, Chang AY, Brack AS, Marshall WF. Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. *PLoS Comput Biol* 2018;**14**:e1005927.
41. Neumann B, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* 2010;**464**:721−7.
42. Pincus Z, Theriot JA. Comparison of quantitative methods for cell-shape analysis. *J Microsc* 2007;**227**:140−56.
43. Yan BP, Lai WHS, Chan CKY, et al. High-throughput, contact-free detection of atrial fibrillation from video with deep learning. *JAMA Cardiol* 2020;5 (1):105−107. Available from: https://doi.org/10.1001/jamacardio.2019.4004.
44. Wu H-Y, et al. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans Graph* 2012;**31**:1−8.
45. Elgharib M, Hefeeda M, Durand F, Freeman WT. Video magnification in presence of large motions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 2015;4119−27.
46. Chen L-C, Papandreou G, Schroff F, Adam H. *Rethinking atrous convolution for semantic image segmentation. arXiv [cs.CV]*. 2017.
47. Barron JL, Fleet DJ, Beauchemin SS, Burkitt TA. Performance of optical flow techniques. In: *Proceedings 1992 IEEE computer society conference on computer vision and pattern recognition*, 1992. Available from: https://doi.org/10.1109/cvpr.1992.223269.
48. Loehr LR, Rosamond WD, Chang PP, Folsom AR, Chambless LE. Heart failure incidence and survival (from the atherosclerosis risk in communities study). *Am. J. Cardiol.* 2008;**101**:1016−22.
49. Pellikka PA, et al. Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. *JAMA Netw Open* 2018;**1**:e181456.
50. Farsalinos KE, et al. Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the EACVI/ASE inter-vendor comparison study. *J Am Soc Echocardiogr* 2015;**28**:1171−81 e2.