Deep learning applications in visual data for benign and malignant hematologic conditions: a systematic review and visual glossary

Andrew Srisuwananukorn,[1] Mohamed E Salama[2] and Alexander T. Pearson[3]

[1]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [2]Sonic Healthcare USA, Austin, TX and [3]Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, IL, USA

Correspondence: A.T. Pearson
apearson5@medicine.bsd.uchicago.edu

Received: June 3, 2022.
Accepted: January 18, 2023.
Early view: January 26, 2023.

https://doi.org/10.3324/haematol.2021.280209

Abstract

Deep learning (DL) is a subdomain of artificial intelligence algorithms capable of automatically evaluating subtle graphical features to make highly accurate predictions, which was recently popularized in multiple imaging-related tasks. Because of its capabilities to analyze medical imaging such as radiology scans and digitized pathology specimens, DL has significant clinical potential as a diagnostic or prognostic tool. Coupled with rapidly increasing quantities of digital medical data, numerous novel research questions and clinical applications of DL within medicine have already been explored. Similarly, DL research and applications within hematology are rapidly emerging, although these are still largely in their infancy. Given the exponential rise of DL research for hematologic conditions, it is essential for the practising hematologist to be familiar with the broad concepts and pitfalls related to these new computational techniques. This narrative review provides a visual glossary for key deep learning principles, as well as a systematic review of published investigations within malignant and non-malignant hematologic conditions, organized by the different phases of clinical care. In order to assist the unfamiliar reader, this review highlights key portions of current literature and summarizes important considerations for the critical understanding of deep learning development and implementations in clinical practice.

Introduction

Recent advances in large-scale data storage, availability, and computational power have led to significant interest in the development of new techniques for "big data" analysis. Rapidly evolving artificial intelligence (AI) algorithms aim to efficiently utilize vast amounts of information with minimal human interaction to address tasks that automate or improve upon human-level assessment. Artificial intelligence takes many forms and includes domains of deep learning (DL), convolutional neural networks (CNN), and other related techniques that are capable of processing imaging data quickly and automatically. Research divisions within commercially successful technology companies have popularized DL models for vision-related tasks, such as facial recognition, image segmentation, object detection, and many other examples that are currently being integrated into daily life.

Within the medical field, visual assessment of digitized clinical imaging and biospecimens by physicians is critical in numerous phases of clinical care for patients. As a result, early investigations that employ clinical DL using histology slides or radiological images within medicine have produced promising results, including diagnosis detection,[1] clinical subtyping,[2] cancer mutation prediction,[3,4] and survival.[5] Recognizing the clinical importance of these algorithms, the US Food and Drug Administration has approved a number of novel AI and DL products.[6]

However, DL algorithms exploring malignant and non-malignant hematologic conditions are still scarce. With digitization tools generating larger biospecimen image databases[7,8] and researchers becoming increasingly familiar with DL techniques, examples of applications in hematology are growing exponentially.[9-14] As such, it is inevitable for hematologists to be familiar with the broad concepts, applications, and limitations of clinical DL.

In this structured narrative review, we aim to describe the general concepts, provide a visual glossary for key terms within image-based DL, and conduct a systematic review to provide an up-to-date assessment of the application of image-based DL in benign and malignant hematology across various phases of patient care.

Neural networks and deep learning

The concept of “deep learning” is poorly defined, imprecise, and often used interchangeably with terms such as “machine learning” and “artificial intelligence.” Traditionally, “artificial intelligence” is the use of automated systems to perform a particular task. “Machine learning” represents a subset of AI in which rules are not explicitly predetermined, but are acquired by training and optimizing parameters based on observed data. Machine learning workflows traditionally separate data into training, validation, and external testing cohorts for model assessment. Examples of machine learning that are probably the most familiar include linear regression, logistic regression, or Cox proportional hazards models. “Deep learning” is a recently-popularized subset of machine learning utilizing a specialized neural-network architecture undergoing millions of arithmetic operations (Figure 1).[15,16] DL architectures are loosely modeled after the complex neural connections of the human brain.[17] Although the term “deep learning” is derived from “deep convolutional neural networks” and has gained interest particularly in clinical research, the strict definition has become increasingly ambiguous and may not completely represent modern state-of-the-art techniques. The field of DL and the list of essential glossary terms are rapidly changing, but in keeping with contemporary clinical manuscripts, this review will use the term “deep learning” to mean “deep convolutional neural networks and other contemporary techniques related to computer vision”. There are also non-image-based neural networks and image-based machine learning architectures without neural networks; however, both are beyond the scope of this current review.

Image preprocessing

A standard workflow[3,18,19] for DL research typically requires preprocessing input images, which can expedite DL training time or improve performance. Preprocessing steps are typically dependent on the modality of the

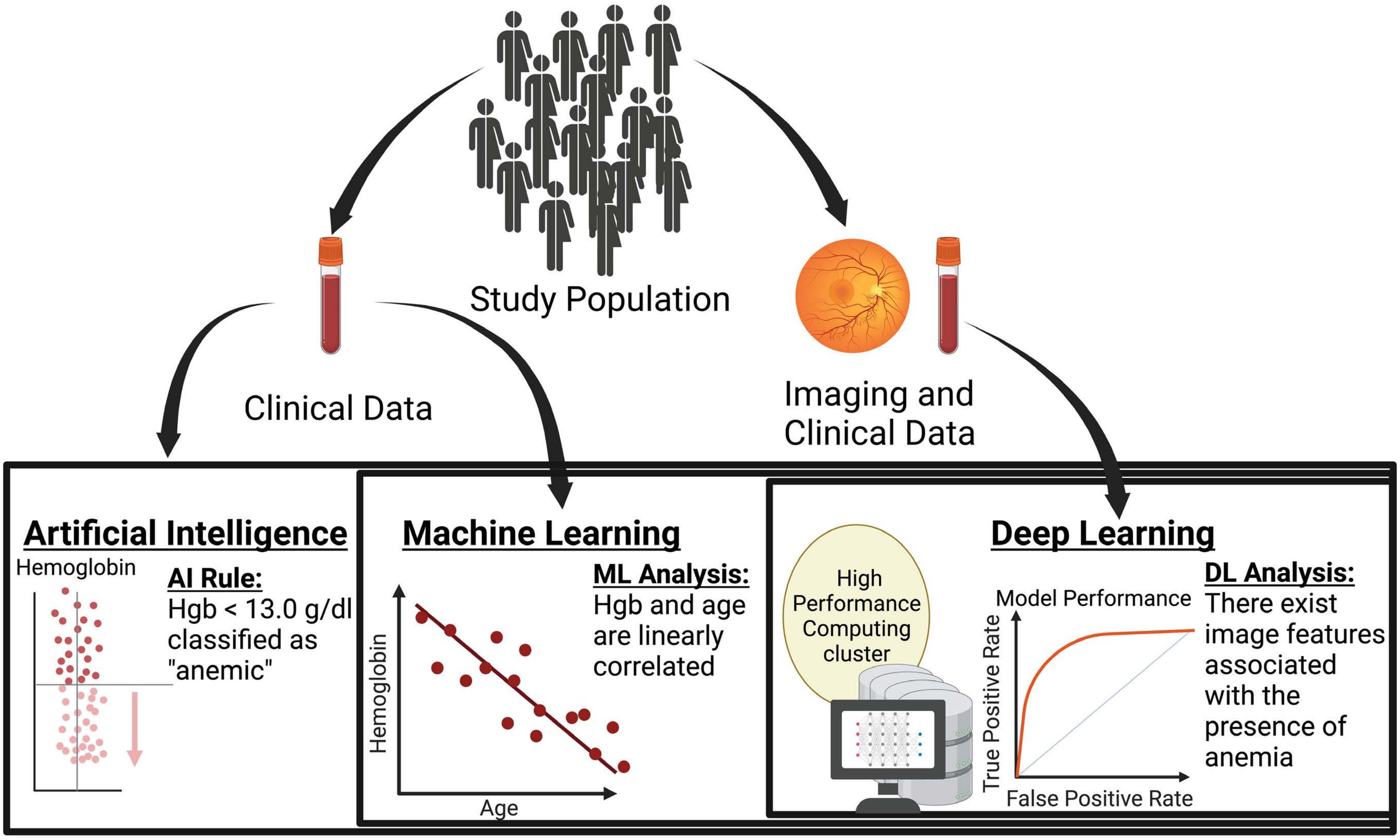
Study Population
Clinical Data
Imaging and Clinical Data
Artificial Intelligence
Hemoglobin
AI Rule:
Hgb < 13.0 g/dl classified as "anemic"
Machine Learning
ML Analysis:
Hgb and age are linearly correlated
Hemoglobin
Age
Deep Learning
High Performance Computing cluster
Model Performance
True Positive Rate
False Positive Rate
DL Analysis:
There exist image features associated with the presence of anemia

Figure 1. Exemplifying differences in “Artificial Intelligence”, “Machine Learning”, and “Deep Learning” with regards to anemia. “Artificial Intelligence” (AI) involves automation of tasks, and can be an explicitly programmed rule to categorize based on the level of a laboratory value. “Machine Learning” (ML) methods, such as linear or logistic regression, derive associations from given training data. More complex image-based “Deep Learning” (DL) models utilize complex architectures termed “Neural Networks” to associate subtle features associated with particular outcomes of interest by using input training data, similar to other machine learning frameworks. In this figure, the clinical condition of anemia based on hemoglobin (Hb) values is used as an example for the above computational frameworks. DL methods may extract features in non-traditional images, such as fundoscopic exams, to derive clinically meaningful categorizations.

image type. While radiological images may be input either whole or with particular Regions of Interest (ROI) segmented, histopathology slides are typically tessellated into smaller tiles representing tissue or segmented cells of interest prior to inclusion into the model. Normalization of pathology images may reduce artifacts specific to a clinical site or particular scanning device, but there is no current standard normalization process. Data augmentation may be performed with random image rotations, vertical or horizontal flips, and simulated compression artifacts to increase the size of the training set and broaden generalizability. In addition to using images alone, researchers can include other data modalities such as clinical information with multi-modal models to supplement image inputs in attempts to improve model performance.

General neural network structure

In a simplified viewpoint of neural network structures (Figure 2), the input image is transformed at various intermediate states, termed "nodes," with each node representing a different graphical feature of the image. As the image is passed from node to node, the connection between each node involves mathematical transformations to represent more complex features in later nodes. Each node can be connected to multiple subsequent nodes simultaneously, and the group of nodes with similar numbers of sequential connections from the input image represent a layer of intermediate nodes. Shallow and Deep neural networks refer to the number of node layers within a particular architecture, but there is no strict definition to differentiate the two. In addition, nodes may not necessarily connect to the nodes in the immediately subsequent layer, but may connect by "skip connections" to nodes in later layers. The penultimate layer of nodes, each representing only a single numerical value, is termed the Logit Layer, the values of which are then normalized between the range of 0 and 1 to give the final probabilities for the outcomes of interest. Common outcomes of interest and examples include object detection, segmentation, classification, regression, survival analysis, and detail optimization (Figure 3).

Information propagation and parameter training

To develop a neural network model, the input image is represented numerically by each pixel. The numerical information is propagated though intermediate nodes and layers towards the direction of the output layer. The connections between nodes are mathematically represented by either non-linear operations or matrix multiplication and addition with potentially millions of trainable parameters, whose values are updated while optimizing the end outcome.

Upon initial model evaluation, the sequential movement of information from input image towards the outcome of interest is deemed "forward propagation" or "forward pass" (Figure 2A-F). To complement an initial prediction, the user defines a particular loss function to quantitatively describe the incorrectness of the model's prediction from available ground truth. Using an additional user-defined optimizer algorithm, the trainable parameters are iteratively adjusted to decrease the loss value in subsequent forward passes. This framework of optimizing parameters in earlier layers using information from the predicted outcome is deemed "back propagation." During training, forward and back propagation are repeated for a defined number of repetitions, or epochs, but training can also be stopped if other defined optimal conditions are met. Given the need for at least 109 calculations per forward pass, parallel computing often requires specific hardware such as Graphics Processing Units (GPU) to expedite necessary matrix operations to be finished within reasonable timeframes.

Convolutions

At the time of writing, the most popular type of DL architecture is the convolutional neural network (CNN). The prototypical CNN algorithm assesses a smaller grid-like portion of each input image prior to propagation towards the next layer (Figure 2C). CNN utilize the convolution operation between layers, which involves matrix multiplication across overlapping sub-sections of the input image to produce a lower-dimensional output representation.

Pre-trained networks and transfer learning

Initially, CNN trained to perform object detection required millions of manually-annotated images, training for days or weeks on industry-grade computational equipment.[20] After training is complete, CNN have traditionally been understood to learn "low-level" general features such as lines, edges, and shapes in earlier layers of the network, but more complex "high-level" features such as faces, patterns, and spatial distributions are learned in subsequent layers that are more closely associated with the evaluated outcome.[21]

In clinical research, it is rare for clinicians to have the resources to develop new CNN architectures with initially random parameters; such a feat requires large-scale databases with expert-level annotations and access to industry-grade supercomputers. Researchers have taken advantage of the learned features along progressive layers by using models previously trained on large databases for non-clinical tasks, but repurposing the final few layers to predict specific clinically-relevant outcomes. The concept of transfer learning involves utilizing a pre-trained network such as those already trained on the ImageNet database of over 1 million general images,[22] initializing the model with the parameters that learned "low-level" features from images unrelated to the application of interest, and allowing the model to retrain and modify parameters in the last few layers to learn “higher-level” features on images for specific patient-related tasks. By utilizing transfer learning, the minimum required dataset and computational power is significantly less than fully training a network from completely random parameters.[23]

Specific deep learning architectures in clinical research

While DL is a framework of neural networks for outcome prediction, each specific model architecture incorporates drastically different complexities with regards to number of layers, connections between layers, functions, and many other highly-engineered features. In fact, newer contempor-

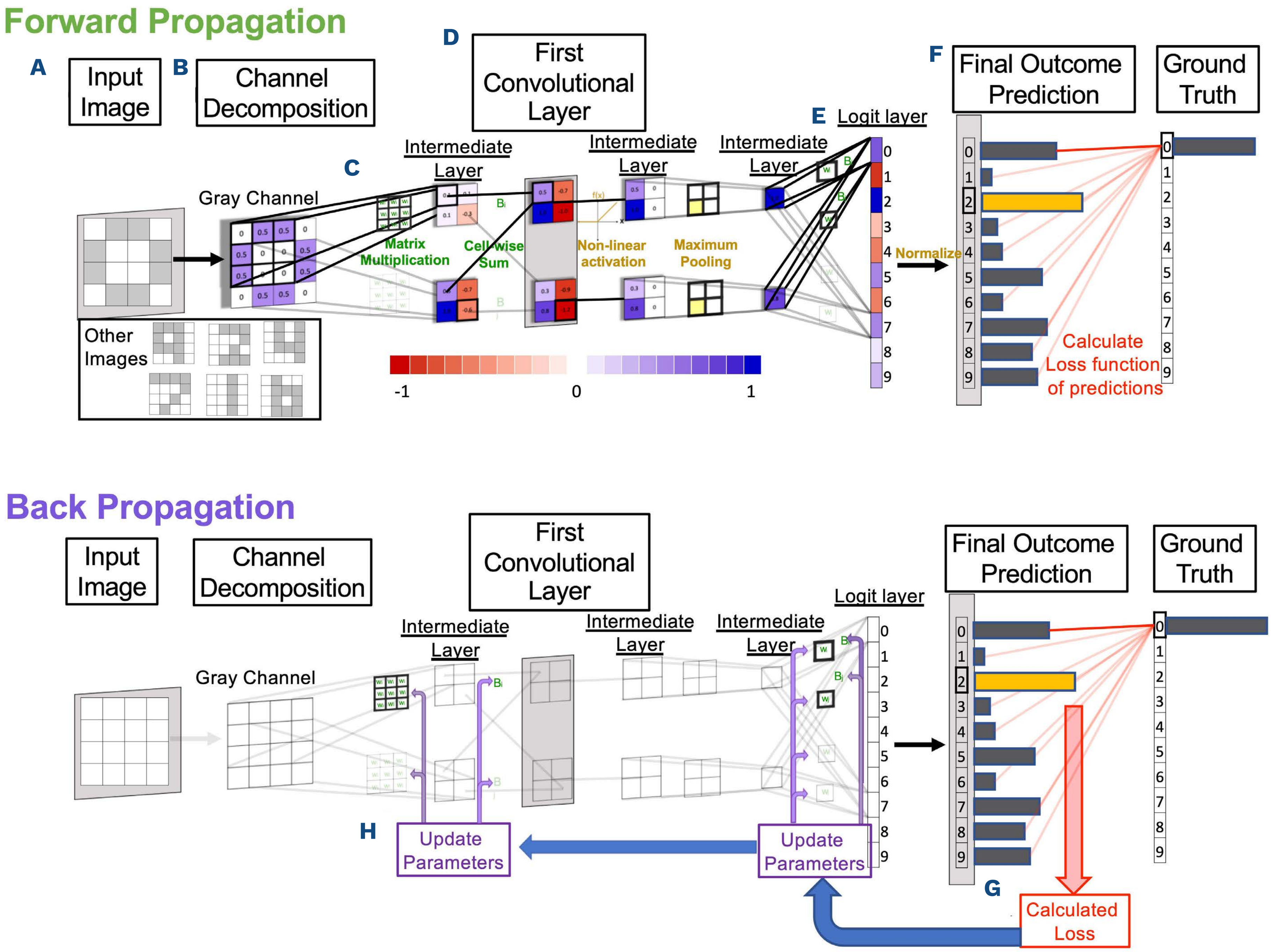

Figure 2. Brief representation of the structure and training phases of deep convolutional neural networks. Collectively, A-H represent “forward propagation” and G-H represent “back propagation.” (A) Images are first passed into the network to predict an outcome of interest. In this example, 4x4 pixelated images of written numbers are used to train a network to predict the numeric value of the image. (B) Pixels are initially converted into numeric values based on pixel intensity. (C) Smaller subsections of the input images are transformed with the convolutional operation, which involves matrix multiplication and addition with trainable parameters. (D) As information is passed into subsequent layers, the image undergoes non-linear transformations, such as the Rectified Linear Unit function that allow the model to represent non-linear relationships within the data. (E) Within this figure, intermediate layers are restructured to a layer of single numerical values in the Logit layer. (F) After propagation through the pre-defined number of convolutional layers, the final activation function normalizes the Logit layer into a distribution of probabilities across the space of available outcomes. The value with the highest probability is deemed the model’s prediction. (A-F) The framework outlined as information is passed from image input to model prediction is termed “forward pass” or “forward propagation.” (G) After the first pass of the model’s predictions, a loss function specific to the outcome data type is calculated to quantitatively assess the level of error produced by the initial prediction. The loss function is chosen before training by the user. Common examples of loss functions are “cross entropy” for categorical outcomes and “mean square error” for regression outcomes. (H) Optimization algorithms iteratively alter the trainable parameters within each of the previous convolutional layers based on the defined loss function. The direction and magnitude of parameter adjustments is calculated by either maximizing or minimizing the loss function in future forward passes, as chosen by the user for the outcome of interest. (G and H) The framework outlined for automatically adjusting earlier parameters to optimize model performance is termed “back propagation.” The process of forward (A-F) and back (G-H) propagation is repeated until a pre-specified set of conditions is fulfilled, typically leading to more accurate predictions.

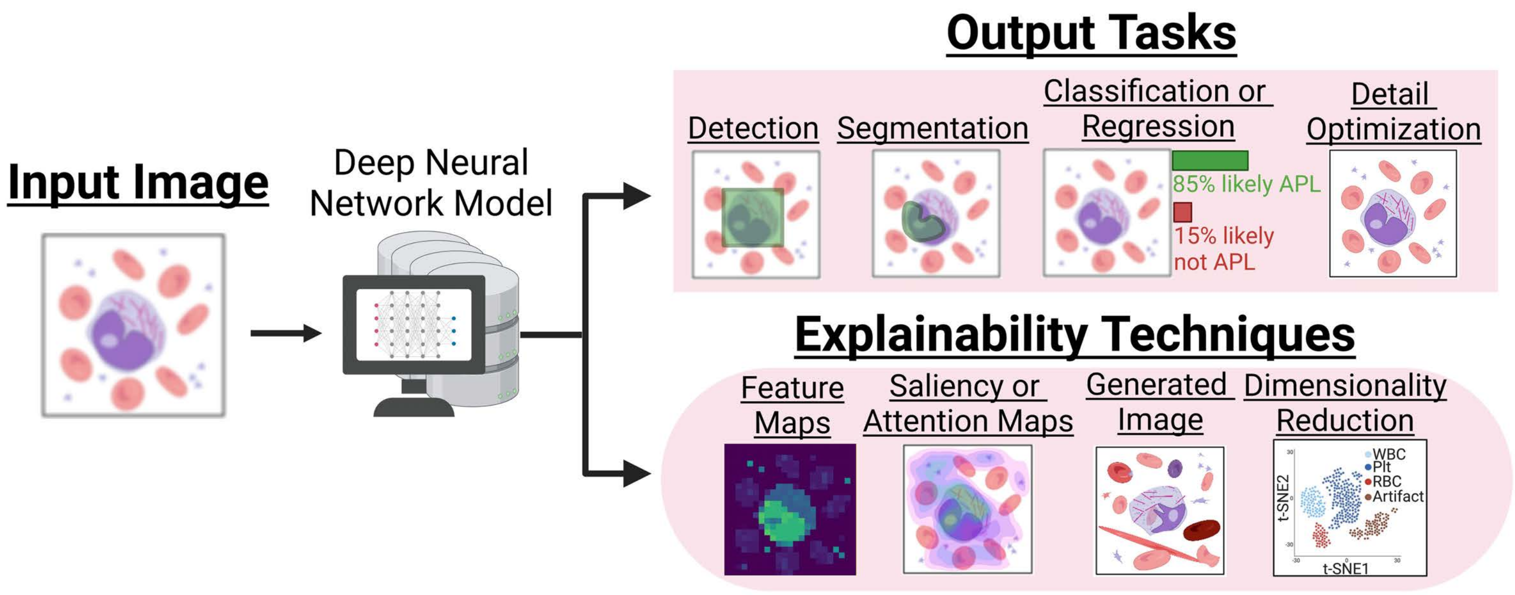

Figure 3. Examples of outcome tasks and explainability methods in Deep Learning. In this example, the initial input image is a promyelocyte with visible Auer Rods as seen on a peripheral blood smear, a possible pathognomonic finding for acute promyelocytic leukemia. Output tasks can include localization of white blood cells (Detection), creation of a region of interest around the nucleus (Segmentation), disease prediction (Classification/Regression), or increasing the visual quality of the input image (Detail Optimization). Explainability methods are necessary to ensure biological feasibility. In exploratory analyses, parameters within the intermediate layers can be directly visualized (Feature Maps), heatmaps can be generated to highlight specific areas associated with the outcome (Saliency or Attention Maps), synthetic images can be generated from noise to represent an outcome of interest (generative adversarial networks), or cluster analyses can be performed with dimensionality reduction techniques.

ary models lack any convolutional layers, and infer local and global image features by other methods.[24]

Thus far, the predominant architecture for hematology-specific questions tend to be from a class of CNN known as Residual Neural Networks (ResNets), which utilize skip connections. Most specific ResNet architectures, such as Inception, EfficientNets, MobileNets, and other various ResNet models are open-source and widely available.[25]

Certain model architectures are engineered to provide an output that is an additional image; these model structures are needed for dimensionality reduction, bounding-box detection, segmentation, and noise reduction tasks. One specific architecture, Autoencoders, are networks that pass an input image through an intermediate lower-dimensional representation, followed by upsizing to a higher-dimensional space to recreate the input image.[26] Theoretically, the lower-dimensional intermediate representation still retains features of the original image which may be clinically or biologically relevant. Similar architectures such as U-Net require additional training data, such as object ROI or low-/high-quality image pairs, to accomplish tasks such as image segmentation or digital optimization.

An additional relevant DL framework utilized is Multiple Instance Learning (MIL)[27,28] and its attention-based derivatives[29,30] including the Clustering-constrained Attention Multiple Instance Learning (CLAM).[31] The main distinction in MIL frameworks is the prediction for data subsets and not for single instances. Specifically, input images are separated into smaller subsets. The entirety of the subset is predicted "positive" if at least one image in the subset is predicted "positive". As an example of MIL in histopathology, a biopsy whole-slide image would be predicted "cancerous" when one extracted tile is predicted as such.[1] This framework may be particularly helpful when single annotations are provided across an entire image, or "weak supervision", and not necessarily labels for each specific segmented ROI. In addition, Attention, or a numeric weight, can be assigned to each image tile to produce weighted predictions, as well as provide explainable heatmaps. Using Attention, CLAM was developed to increase the speed of MIL and reduce the noise from irrelevant image tiles.

Vision Transformers (ViT) are a novel technique that do not utilize the convolution operator.[24] The entire image is separated into a grid of sub-images that are analyzed in parallel along with the relative location of each sub-image. With this method, global relationships across the entire image may be learned by the model as opposed to only local features that are seen by the previously-described CNN.

Currently, most architectures for hematology-specific questions utilize ResNet architectures, with just a few examples also incorporating MIL. However, the emergence of ViT and CLAM frameworks are part of a changing landscape of implemented DL architectures. In general, the choice of model architecture is somewhat informed by expected outcome task, but it is still largely empiric. However, there are broad advantages and disadvantages for each of the previously-mentioned frameworks. With weak supervision, MIL tends to require significantly larger amounts of training data than ResNets.[1] CNN and ViT perform equally well at the scale of currently available clinical datasets. However, ViT are superior to CNN for larger scale datasets and are more computationally efficient with significantly fewer parameters.[24] There are numerous methods to attempt to explain the inner mechanisms of

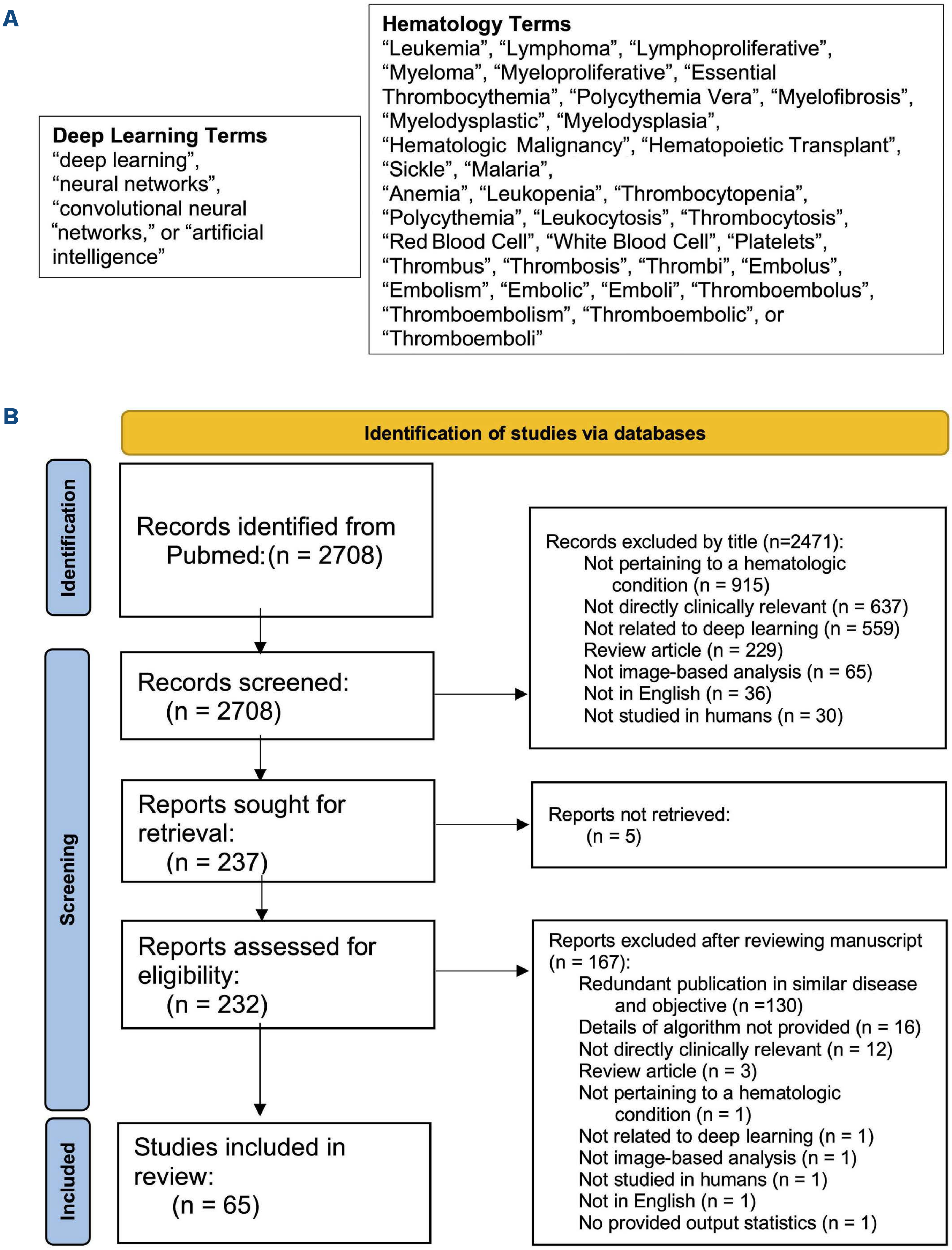
A
Deep Learning Terms
"deep learning",
"neural networks",
"convolutional neural
"networks," or "artificial
intelligence"
Hematology Terms
"Leukemia", "Lymphoma", "Lymphoproliferative",
"Myeloma", "Myeloproliferative", "Essential
Thrombocythemia", "Polycythemia Vera", "Myelofibrosis",
"Myelodysplastic", "Myelodysplasia",
"Hematologic Malignancy", "Hematopoietic Transplant",
"Sickle", "Malaria",
"Anemia", "Leukopenia", "Thrombocytopenia",
"Polycythemia", "Leukocytosis", "Thrombocytosis",
"Red Blood Cell", "White Blood Cell", "Platelets",
"Thrombus", "Thrombosis", "Thrombi", "Embolus",
"Embolism", "Embolic", "Emboli", "Thromboembolus",
"Thromboembolism", "Thromboembolic", or
"Thromboemboli"
B
Identification of studies via databases
Identification
Screening
Included
Records identified from Pubmed: (n = 2708)
Records screened: (n = 2708)
Records excluded by title (n=2471):
Not pertaining to a hematologic condition (n = 915)
Not directly clinically relevant (n = 637)
Not related to deep learning (n = 559)
Review article (n = 229)
Not image-based analysis (n = 65)
Not in English (n = 36)
Not studied in humans (n = 30)
Reports sought for retrieval: (n = 237)
Reports not retrieved: (n = 5)
Reports assessed for eligibility: (n = 232)
Reports excluded after reviewing manuscript (n = 167):
Redundant publication in similar disease and objective (n =130)
Details of algorithm not provided (n = 16)
Not directly clinically relevant (n = 12)
Review article (n = 3)
Not pertaining to a hematologic condition (n = 1)
Not related to deep learning (n = 1)
Not image-based analysis (n = 1)
Not studied in humans (n = 1)
Not in English (n = 1)
No provided output statistics (n = 1)
Studies included in review: (n = 65)

Figure 4. Literature search. (A) Search terms to extract relevant manuscripts related to deep learning in malignant and non-malignant hematology. Articles were queried in PubMed using one "Deep Learning" term in addition to one "Hematology" term. (B) PRISMA diagram of "Deep Learning in Hematology" survey. Initially 2,708 articles were found from a PubMed query. After initial review of abstracts and article titles, 237 reports were deemed eligible for further review of full manuscripts. Finally, 65 articles were included for the current narrative review. Justification for exclusion are provided.

standard CNN,[32] but similar methods to "open the black box" of ViT are currently under development.[33]

Explainability

While "explainability" in DL research is loosely defined, in this review, "explainability" refers to the efforts in describing DL models and predictions in humanly-understandable concepts.[34]

Although DL may empirically exhibit a high performance, DL is often criticized for its highly complex mechanisms and is often thought of as a "black box." In multiple examples, seemingly high-performing models often utilize artifact or contextually irrelevant features for its predictions, as the artifactual features may be unintentionally over-represented in certain imaging subgroups.[35] Multiple methods are under development to explain and validate biologically reasonable predictions. As such, explainability is increasingly important in clinical AI development and in developing physicians' trust of DL.[36]

To give just a few examples, unsupervised data dimensionality reduction methods such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are statistical techniques used to group visually similar input images into clusters, which may overlap with relevant outcomes. These methods are also popularized in non-imaging data such as single cell molecular and cytometry time-of-flight analyses. Feature maps are direct visual representations of the intermediate trained parameters. Plotting Attention scores or using Saliency map methods such as Grad-CAM or Smooth-Grad can overlay heat-maps upon the input image to highlight relevant visual cues associated with the outcome of interest.[35] For example, the heatmap explainability methods of a peripheral blood smear image may highlight pathognomonic Auer Rods for the accurate diagnosis of acute promyelocytic leukemia (Figure 3). More complex methods such as Generative Adversarial Networks are architectures trained to generate synthetic images, which can create representations of a particular class or outcome.[37]

Metrics

Common performance metrics for the evaluation of DL classification models include Area Under the Receiver Operator Curve (AUROC), sensitivity, specificity, and accuracy. The AUROC represent the tradeoff between true and false positive rates for a binary model along a range of possible threshold values. AUROC values nearing 1.0 represent a model with perfect discriminatory power, and values tending towards 0.5 perform no better than random chance.

For segmentation tasks, the Sørensen-Dice similarity coefficient (Dice) represents the overlap between the predicted area of interest with the ground truth, where a Dice coefficient of 1.0 represents ideal predictive overlap. Other segmentation metrics include the similarly defined Jaccard index, also known as Intersection over Union (IoU).

Literature review for clinical application of deep learning in hematologic conditions

A Boolean query was submitted to PubMed to extract articles created between January 1, 1990, and August 1, 2022. Search terms included both a “deep learning” and a “hematology” specific term (Figure 4A). The query resulted in 2,708 initial articles. Further refinement by manual review by one author excluded a large number of articles (Figure 4B), resulting in 65 manuscripts. General trends and findings of the resulting articles are described in the context of how DL has been utilized to enhance phases of clinical care within various hematologic conditions, including task automation, detail optimization, disease detection, differential diagnosis, disease classification, risk prediction, complication assessment, therapy response, and survival prediction (Figure 5). General considerations for critical appraisal of the following manuscripts include performance metrics, use of external or prospective validation cohorts, use of explainability

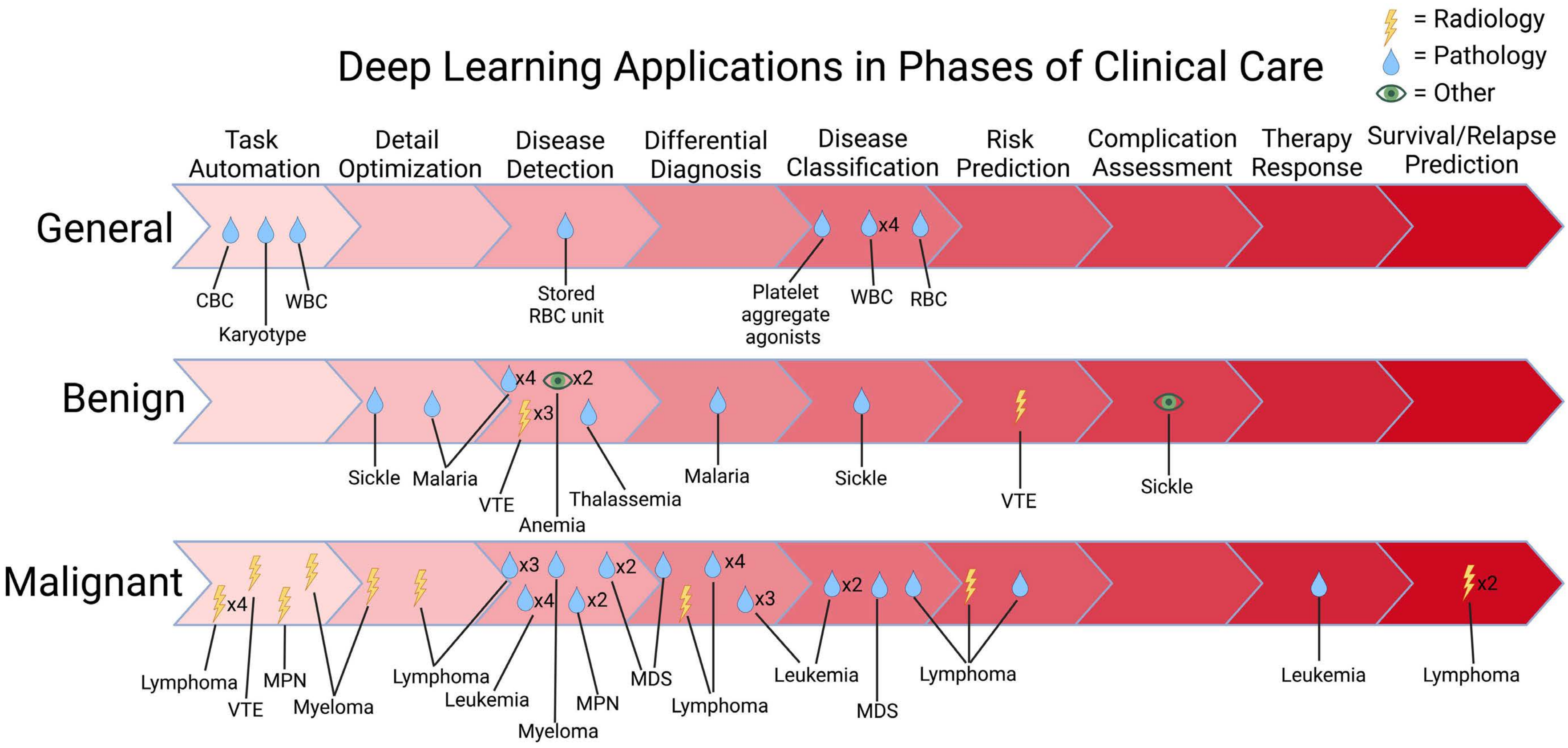

Figure 5. Deep Learning applications within 65 malignant and non-malignant hematology manuscripts. Applications are divided into separate phases of clinical care, including task automation, detail optimization, disease detection, differential diagnosis, disease classification, risk prediction, complication assessment, therapy response, and survival/relapse prediction. Image domains include radiological, pathological, and other atypical image types such as electrocardiograms or funduscopic exams. Specific image modalities are detailed in Tables 1-4. CBC: complete blood count; MDS: myelodysplastic syndromes; MPN: myeloproliferative neoplasms; RBC: red blood cell; VTE: venous thromboembolism; WBC: white blood cell.

Table 1. Summary and performance metrics of deep learning applications in hematology for task automation and detail optimization.

Author	Disease	Clinical task	Image modality	Total patients	Total images (N)	Main result	Value	Validation strategy	Explainability strategy	Human comparison
Task automation										
Fan[41]	General	Localize and segment WBC	PBS	-	925	Dice	0.97-0.98	Internal	-	No
Alam[39]	General	Automate cell count	PBS	-	364	Accuracy	0.80-0.95	Internal	-	No
Vajen[40]	General	Identify and rotate chromosomes	Karyogram	-	330,131	Accuracy	0.99	Internal	-	No
Jemaa[41]	Lymphoma	Localize NHL lesions	PET-CT	1,695	3,664	Dice	0.87	Internal	-	No
Weisman[44]	Lymphoma	Localize pediatric HL lesions	PET-CT	100	-	Dice	0.86	Internal	-	Yes
Weisman[43]	Lymphoma	Localize lymph nodes in HL and DLBCL	PET-CT	90	-	TPR	0.85	Internal	-	Yes
Sadik[45]	Lymphoma	Localize FDG avid lesions in DLBCL	PET-CT	153	-	Human agreement	0.81	Internal	-	Yes
Yang[46]	MPN	Localize spleen and calculate volume	CT	-	138	Dice	0.95	Internal	-	Yes
Xu[42]	Myeloma	Localize myeloma bone lesions	PET-CT	12	-	Dice	0.73	Internal	-	No
Liu[47]	PE	Localize and segment PE to calculate clot burden	CT	878	878	AUROC	0.93	External	-	No
Detail optimization										
Theruvath[51]	Lymphoma	Reduce noise in lymphoma images	PET-MRI	-	-	Proposed dose reduction	0.5	External	-	No
Shaw[49]	Malaria	Digitally enhance peripheral blood images	PBS	-	74	Δ Absolute variance	0.25	Internal	-	No
Huber[50]	Myeloma	Reduce noise in myeloma images	CT	10	-	Proposed dose reduction	0.25	Internal	-	No
De Haan[48]	Sickle	Digitally enhance mobile-device photos	PBS	96	96	AUROC	1	Internal	-	No

MPN: myeloproliferative neoplasm; PE: pulmonary embolus; WBC: white blood cell; NHL: non-Hodgkin lymphoma; HL: Hodgkin lymphoma; DLBCL: diffuse large B-cell lymphoma; FDG: fluorodeoxyglucose; PBS: peripheral blood smear; PET-CT: positron emission tomography / computed tomography; PET-MRI: positron emission tomography / magnetic resonance imaging; Dice: Sørensen-Dice similarity coefficient; TPR: true positive rate; AUROC: Area Under Receiver Operator Curve; N: number.

Table 2. Summary and performance metrics of deep learning applications in hematology for disease detection.

Author	Disease	Clinical task	Image modality	Total patients (N)	Total images	Main result	Value	Validation strategy	Explainability strategy	Human comparison
Disease detection										
Mitani[74]	Anemia	Detect anemia	Fundoscopy	57,163	114,205	AUROC	0.88	Internal	Saliency map	No
Kwon[73]	Anemia	Detect anemia	ECG	44,537	70,074	AUROC	0.9	External	Saliency map	No
Lee[57]	Anemia	Detect HbH inclusions	PBS	110	515	AUROC	0.84	Internal	-	No
Kainz[72]	DVT	Detect DVT	Ultrasound	255	-	AUROC	0.77-0.87	Prospective	-	No
Doan[58]	General	Classify stored RBC quality	IFC	38 RBC units	40,900	Human agreement	0.77	External	UMAP, t-SNE	Yes
Shafique[62]	Leukemia	Detect ALL by FAB subtype	PBS	-	454	Accuracy	0.96-0.99	Internal	-	No
Sahlol[61]	Leukemia	Detect ALL	PBS	76	10,661	Accuracy	0.83-0.96	Internal	-	No
Sidhom[12]	Leukemia	Detect APL	PBS	106	5,547	AUROC	0.86	Internal	Saliency map, UMAP	Yes
Eckardt[13]	Leukemia	Detect APL	BMA	1,335	-	AUROC	0.86	Internal	Saliency map	No
Syrykh[67]	Lymphoma	Detect FL	LNB	-	443	AUROC	0.63-0.69	External	-	No
Li[66]	Lymphoma	Detect DLBCL	LNB	-	4,665	Accuracy	0.91	External	-	No
Sibille[68]	Lymphoma	Detect DLBCL	PET-CT	629	-	AUROC	0.95	Internal	-	No
Zhou[69]	Lymphoma	Detect MCL	PET-CT	142	-	Sensitivity	0.84	External	-	No
Rajaraman[52]	Malaria	Detect malaria	PBS	200	13,779	AUROC	0.99	Internal	-	No
Rajaraman[53]	Malaria	Detect malaria	PBS	200	-	AUROC	0.99	Internal	Saliency map	No
Kuo[54]	Malaria	Detect malaria	PBS	-	36	AUROC	1	Internal	-	Yes
Li[56]	Malaria	Detect malaria and Babesia	PBS	-	21,236	Accuracy	0.95-0.99	External	t-SNE	No
Mori[63]	MDS	Detect dysplastic neutrophils	BMA	-	35	AUROC	0.94	Internal	-	No
Acevedo[59]	MDS	Detect dysplastic neutrophils	PBS	144	249	AUROC	0.98	Internal	t-SNE, Saliency map	No
Sirinukunwattana[64]	MPN	Detect MPN	BMB	-	131	AUROC	0.98	Internal	PCA	No
Kimura[60]	MPN	Detect MPN	PBS	234	344	AUROC	0.97-0.99	Internal	-	No
Gehlot[65]	Myeloma	Detect multiple myeloma	BMA	72	74,996	AUROC	0.98	Internal	t-SNE	No
Huang[70]	PE	Detect PE	CT	1,971	1,997	AUROC	0.85	External	Saliency map	No
Huang[71]	PE	Detect PE using multimodal network	CT	1,794	1,837	AUROC	0.95	Internal	-	No

DVT: deep vein thrombosis; MDS: myelodysplastic syndrome; MPN: myeloproliferative neoplasm; PE: pulmonary embolus; HbH: hemoglobin H (α-thalassemia); RBC: red blood cell; ALL: acute lymphoblastic leukemia; FAB: French-American-British classification; APL: acute promyelocytic leukemia; FL: follicular lymphoma; DLBCL: diffuse large B-cell lymphoma; MCL: mantle cell lymphoma; ECG: electrocardiogram; PBS: peripheral blood smear; IFC: imaging flow cytometry; BMA: bone marrow aspirate; LNB: lymph node biopsy; PET-CT: positron emission tomography / computed tomography; BMB: bone marrow biopsy; CT: computed tomography; AUROC: Area Under Receiver Operator Curve; UMAP: Uniform Manifold Approximation and Projection; t-SNE: t-distributed Stochastic Neighbor Embedding; PCA: Principal Component Analysis; N: number.

Table 3. Summary and performance metrics of deep learning applications in hematology for differential diagnosis and disease classification.

Author	Disease	Clinical task	Image modality	Total patients	Total images (N)	Main result	Value	Validation strategy	Explainability strategy	Human comparison
Differential diagnosis										
Ahmed[77]	Leukemia	Differentiate between AML, ALL, CML, and CLL	PBS	-	903	Accuracy	0.82	Internal	-	No
Huang[78]	Leukemia	Differentiate between AML, ALL, and CML	BMA	104	104	Accuracy	0.95	Internal	-	No
Schouten[9]	Leukemia	Differentiate ALL lymphoblasts from WBC	PBS	-	250	AUROC	0.97	Internal	Saliency map, t-SNE	No
Achi[79]	Lymphoma	Differentiate between DLBCL, BL, and SLL	LNB	-	128	Accuracy	1	Internal	-	No
Guan[80]	Lymphoma	Differentiate NHL from other cancers	FNA	-	80	Accuracy	0.81	Internal	-	No
Mohlman[81]	Lymphoma	Differentiate between BL and DLBCL	LNB	70	10,818	AUROC	0.92	Internal	-	No
Miyoshi[82]	Lymphoma	Differentiate between DLBCL, FL, and lymph	LNB	-	388	AUROC	0.99-1.0	Internal	-	Yes
Yun[83]	Lymphoma	Differentiate PCNSL from GBM	MRI		195	AUROC	0.49	External	-	Yes
Li[75]	Malaria	Differentiate between infectious RBC inclusions	PBS	-	24,358	AUROC	1	Internal	t-SNE	Yes
Kimura[76]	MDS	Differentiate between MDS and AA	PBS	1,165	3,261	AUROC	0.99	Internal	t-SNE	No
Disease classification										
Zhao[84]	General	Classify WBC	PBS	-	1,498	Accuracy	0.93	Internal	-	No
Durant[87]	General	Classify RBC	PBS	97	3,737	Accuracy	0.91	Internal	-	No
Lippeveld[85]	General	Classify WBC	IFC	2	98,013	Accuracy	0.7	Internal	UMAP	No
Wu[86]	General	Classify WBC	BMA		122	Sensitivity	0.86	Internal	-	Yes
Zhou[89]	General	Classify platelet aggregates by agonist	IFC	1	60,000	Accuracy	0.77	Internal	t-SNE	No
Matek[11]	General	Classify WBC	BMA	961	-	Sensitivity	0.20-0.91	External	UMAP, Saliency map	No
Rehman[90]	Leukemia	Classify ALL into FAB subtypes	BMA		330	Accuracy	0.9778	Internal	-	No
Eckardt[14]	Leukemia	Classify NPM1 mutation in AML	BMA	1,251	-	AUROC	0.92	Internal	Saliency map	No
Swiderska-Chadaj[92]	Lymphoma	Classify MYC rearrangement in DLBCL	LNB	287	-	Accuracy	0.93	External	-	No
Brück[91]	MDS	Assess diagnosis, risk factors, and genomics in MDS	BMA	205	-	AUROC	0.58 - 0.94	Internal	UMAP	No
Xu[88]	Sickle	Classify sickle and non-sickle RBCs	PBS	8	434	AUROC	0.98	Internal	-	No

MDS: myelodysplastic syndrome; AML: acute myeloid leukemia; ALL: acute lymphoblastic leukemia; CML: chronic myeloid leukemia; CLL: chronic lymphocytic leukemia; WBC: white blood cell; DLBCL: diffuse large B-cell lymphoma; BL: Burkitt's lymphoma; SLL: small lymphocytic lymphoma; NHL: non-Hodgkin lymphoma; FL: follicular lymphoma; PCNSL: primary central nervous system lymphoma; GBM: glioblastoma multiforme; RBC: red blood cell; AA: aplastic anemia; AB: French-American-British classification; PBS: peripheral blood smear; BMA: bone marrow aspirate; LNB: lymph node biopsy; FNA: fine needle aspirate; MRI: magnetic resonance imaging; IFC: imaging flow cytometry; AUROC: Area Under Receiver Operator Curve; t-SNE: t-distributed stochastic neighbor embedding; UMAP: uniform manifold approximation and projection; N: number.

Table 4. Summary and performance metrics of deep learning applications in hematology for advanced phases of patient care, including risk prediction, complication assessment, therapy response, and survival/relapse prediction.

Author	Disease	Clinical task	Image modality	Total patients	Total images (N)	Main result	Value	Validation strategy	Explainability strategy	Human comparison
Risk prediction										
Irshaid[93]	Lymphoma	Predict transformation of CLL or FL to DLBCL	BMB	61	-	AUROC	0.73-0.86	Internal	-	No
Jullien[94]	Lymphoma	Segment and quantify muscle tissue as a prognostic marker in DLBCL	CT	239	-	Dice	0.97	External	-	No
Cahan[95]	PE	Assess severity of PE	CT	363	-	AUROC	0.88	Internal	Saliency map	No
Complication assessment										
Cai[96]	Sickle	Detect retinopathy complication in SCD	UWF-FP	190	1,182	AUROC	0.99	Internal	Saliency map	No
Therapy response										
Doan[97]	Leukemia	Detect residual ALL cells after treatment	IFC	30	-	Accuracy	0.88	Internal	t-SNE	No
Survival/relapse prediction										
Guo[98]	Lymphoma	Predict relapse in ENKTL	PET-CT	84	-	AUROC	0.88	Internal	-	No
Lisson[99]	Lymphoma	Predict relapse in MCL	CT	30	-	AUROC	0.7	Internal	-	No

MDS: myelodysplastic syndrome; PE: pulmonary embolus; CLL: chronic lymphocytic leukemia; FL: follicular lymphoma; DLBCL: diffuse large B-cell lymphoma; MCL: mantle cell lymphoma; SCD: sickle cell disease; ALL: acute lymphoblastic leukemia; ENKTL: extranodal natural killer/T-cell lymphoma; BMB: bone marrow biopsy; CT: computed tomography; BMA: bone marrow aspirate; UWF-FP: ultra-widefield color fundus photographs; IFC: imaging flow cytometry; PET-CT: positron emission tomography / computed tomography; AUROC: Area Under Receiver Operator Curve; Dice: Sørensen-Dice similarity coefficient; UMAP: uniform manifold approximation and projection; t-SNE: t-distributed stochastic neighbor embedding; N: number.

methods, and comparison with human expert performance (Tables 1-4).

Task automation

Routine clinical workflows in pathology and radiology may involve repetitive actions. Automation models can be developed to increase efficiency and decrease physician burden for tasks such as counting cell types in peripheral blood smears or contouring the borders of suspicious lesions on imaging. For pathology workflows, DL models trained to contour white blood cell (WBC) borders in peripheral blood smears were highly effective with near perfect Dice co-efficients in multiple cohorts.[38] Automatic detection of cells can be put through downstream analyses and provide an automated cell count, for which DL-based methods achieve high accuracy.[39]

In addition, chromosomal analyses are standard for diagnosis and prognostication for multiple hematologic malignancies. Manual segmentation and rotation of digital karyograms is time-consuming, but automated models can significantly expedite throughput.[40] In radiology workflows, contouring suspicious lesions or organs can help characterize downstream parameters such as volume, width, and avidity. Hypermetabolic lesions on PET/CT have been localized with DL algorithms for multiple adult and pediatric lymphomas or multiple myeloma lesions.[41,42] Segmentation metrics were reportedly high, with Dice coefficient 0.86-0.98 among various lymphomatous conditions.[43-45] For other conditions, the automated volume calculation of particular regions of interest have been explored in myeloproliferative neoplasms (MPN) for spleen volume,[46] as well as clot burden quantification for new pulmonary emboli.[47]

Detail optimization

For expert diagnosticians, image quality is critical for the identification of disease. Using U-Net architectures, DL-enhanced images may improve user readability and potentially reduce the amount of toxic contrast material given to patients. Enhancement of peripheral blood images to assess red blood cell (RBC) aberrations have yielded promising results. For sickle cell disease, mobile-device photos of peripheral blood have been digitally upscaled to match laboratory microscope quality; upon further validation, the upscaled images retained relevant visual cues with near-perfect classification.[48] However, similar attempts to detect malaria RBC inclusion were less successful, noting that CNN-based enhancement of peripheral blood images was insufficient to resolve parasites that were not already easily distinguishable at low resolution.[49] Multiple optimization efforts in radiology have investigated whether DL can improve image quality from lower-contrast images, which may help spare patients from nephrotoxic or radioactive risks. For both positron emission tomography/magnetic resonance imaging (PET-MRI) in lymphomatous conditions and computed tomography (CT) scans in multiple myeloma, authors have concluded that reduced contrast volumes may be feasible while still maintaining diagnostic quality.[50, 51]

Disease detection

In clinical practice, a common initial diagnostic step for hematologic disorders is the analysis of peripheral blood to observe morphologic abnormalities of RBC, WBC, and platelets. The detection of structural RBC aberrations can identify certain infectious diseases and hemoglobinopathies. In endemic areas of malaria, the *Plasmodium* parasites are often identified by light microscopy as RBC inclusions. Multiple DL initiatives report high accuracy and good model performance for the diagnosis of malaria from peripheral blood in both cross-validated and external cohorts.[52-56] Other RBC aberrations, such as hemoglobin H inclusions in α-thalassemia, can be detected by DL with appropriate peripheral blood staining protocols.[57] With regards to transfusion medicine needs, the quality and degradation of RBC products prior to transfusion can also be determined with DL methods. Using explainability techniques, Doan *et al.* explored their proposed autoencoder network trained on RBC images to identify novel features associated with poor storage quality RBC products.[58]

For certain disorders, the detection of aberrant WBC morphologies from peripheral blood is paramount. DL algorithms consistently detect dysplastic neutrophils pathognomonic for myelodysplastic syndrome (MDS),[59] as well as other white blood precursors to aid in the diagnosis of MPN,[60] acute promyelocytic leukemia (APL),[12] or acute lymphoblastic leukemia (ALL).[61,62] Many DL models for WBC detection have performed with high accuracy and AUROC upon internal validation strategies. If translated into clinical practice, DL models for peripheral blood assessment may expedite critical diagnoses which necessitate emergent therapy, such as APL.

Particularly for myeloid malignancies, bone marrow assessment is usually needed to establish a diagnosis. DL can detect particular cellular morphologies of neutrophils, megakaryocytes, promyelocytes, and plasma cells associated with MDS,[63] MPN,[64] APL,[13] and multiple myeloma,[65] respectively. Similarly for lymphoid malignancies, assessing lymph node architectures can aid the diagnosis of various lymphomas, such as diffuse large B-cell lymphoma (DLBCL)[66] or follicular lymphoma (FL).[67] DL models developed by Li *et al.* maintained high accuracy for the diagnosis of DLBCL from lymph biopsies across four separate institutional cohorts.[66] Furthermore, Syrykh *et al.* utilized the clinical challenge of differentiating follicular lymphoma from follicular hyperplasia to develop a novel DL method quantifying prediction uncertainty, which is not often reported in DL studies. With their uncertainty method, the authors report higher classification capabilities when only considering the newly categorized low-uncertainty images.[67]

In addition to pathologic analysis, clinical guidelines commonly suggest radiologic assessment for the initial workup of suspected malignancy or thrombosis. Using PET/CT images, DL models exhibit high classification of the hypermetabolic lesions for DLBCL diagnosis.[68] However, similar attempts using PET/CT images of mantle cell lymphoma (MCL) patients are challenged with tradeoffs between sensitivity and false positive rates for diagnosis in external cohorts.[69] For select non-malignant conditions, multiple studies explored DL for the expedited and more affordable diagnosis of pulmonary emboli (PE) and deep vein thromboses (DVT), for which a diagnosis may require immediate intervention.[70-72] Huang *et al.* integrated clinical data in conjunction with CT scans to improve their DL model for PE detection. The authors report that multi-modal models exhibit higher classification performance than image-only DL models.[71] In addition, automated detection of common thrombotic conditions may reduce the financial burden, with cost analyses revealing positive financial benefit to health care systems.[72]

Finally, a particularly novel use of DL is the prediction of disease from imaging modalities beyond standard pathologic or radiologic domains. Multiple studies have shown that anemia can be detected with high accuracy utilizing DL on atypical modalities such as electrocardiograms (ECG)[73] or funduscopic examinations.[74] Both authors have implemented explainability methods to reveal features associated with anemia, such as QRS complexes in ECG or optic disk aberrations in funduscopic images. Thus, screening for anemia may offer a low-cost benefit for patients already undergoing these common examinations.

Differential diagnosis

Various hematologic conditions share similar features and presentations, posing challenges in providing a definitive diagnosis in clinical scenarios where radiologic findings may be non-specific and pathological morphologies may be subtle. Differentiating among possible diagnoses is a common clinical task, and various approaches of DL have been explored as a potential means to increase objectivity towards a true diagnosis. For example, Li *et al.* used transfer learning to pre-train their model with images of common household objects, such as bananas, rings, and pears to learn the analogous morphologies of similarly-shaped RBC inclusions of *Toxoplasma*, *Plasmodium*, and *Babesia*.[75]

Interestingly, DL models have been shown to better extract subtle features for disease differentiation than can be assessed by humans. Cytopenias can be a common presentation for either MDS or aplastic anemia (AA) patients. Though either diagnosis typically requires bone marrow biopsy assessment, Kimura *et al.* trained a DL model on peripheral blood images to accurately differentiate between the two conditions.[76] Newly diagnosed leukemia patients commonly present with blasts in peripheral blood. The categorization of blasts into either myeloid or lymphoid lineages requires identifying cell-surface markers by flow cytometry; thus, visualization of blasts is not usually sufficient for classification. Similarly, lymphoma histology share visual commonalities and require immunohistochemical staining of cell-surface markers on biopsy specimens. To address these classification challenges, DL algorithms reportedly differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) utilizing only peripheral blood or bone marrow images,[9,77,78] and similarly among various non-Hodgkin lymphomas (NHL) utilizing standard hematoxylin and eosin (H&E) lymph node biopsy images.[79-82]

For patients with malignant brain lesions found on MRI imaging, clinicians may be tasked to differentiate between primary central nervous system lymphoma (PCNSL) and glioblastoma multiforme (GBM).[83] DL models for this revealed seemingly high initial performance but with a significant reduction to an AUROC of nearly 0.5 in external cohorts.[83] The problem of generalizing results highlights the continued need for critical appraisal of any newly-developed DL model across patient populations.

Disease classification

The classification of blood cells in standard peripheral blood smear review is a ubiquitous task useful in a broad array of diseases. The differential of WBC is necessary to stratify the likelihood of the malignant and non-malignant causes of WBC abnormalities. Numerous studies developed DL models as a single cell WBC classifier. Across the studies, performance remained robust, with the majority of studies achieving accuracies above 90% and explainability techniques highlighting sensitive cellular features.[11,84-86] However, validation upon external cohorts, which commonly reveal a lower performance,[11] is still needed prior to deployment in clinical practice. In addition to WBC classification, the categorization of RBC morphologies is useful within various anemias,[87] including sickle cell disease.[88] To explore platelet abnormalities, Zhou *et al.* developed a highly accurate DL model predicting the identity of agonists causing platelet aggregation using imaging flow cytometry.[89]

Specific sub-classification of diagnoses is often necessary to guide prognostication, counseling, and therapeutic considerations. In numerous non-hematologic applications, previous DL models can accurately further categorize various cancers into genetic and clinical subtypes,[4] which has led to similar explorations within leukemic and lymphomatous conditions. For leukemic classifications, ALL bone marrow images can be separated into the historically relevant French-American-British (FAB) classifications.[90] Furthermore, genomic subtypes may be accurately identified by DL models; Eckardt *et al.* identified *NPM1* mutations among newly diagnosed AML patients and characterized novel cellular morphological features that had not previously been reported.[14] Broader DL efforts to identify each clinically relevant molecular or cytogenetic abnormality have been attempted for MDS sub-classifications.[91] For lymphoma, Swiderska-Chadaj *et al.* developed a DL model predicting *MYC* gene rearrangements in DLBCL patients using lymph node biopsy images. Though *MYC* rearrangement is typically assessed with ancillary fluorescent *in situ* hybridization, the DL model using only H&E images maintained high accuracy upon external cohorts.[92]

Advanced stages of patient care

There are currently few examples of DL for the assistance of later stages of patient care, including risk prediction, complication assessment, therapy response, and survival prediction. For such tasks, the disease processes and image modalities are heterogenous. Risk has been assessed with CT images or digitalized bone marrow biopsies (BMB) for DLBCL outcomes. DL models predict the transformation of low-grade lymphomas to high-grade DLBCL using BMB images,[93] and, furthermore, known clinical risk factors such as sarcopenia can be extracted and quantified in CT images of DLBCL patients.[94] Risk in thrombotic conditions can be characterized automatically using DL classification of right ventricular strain in chest imaging for PE workup.[95] Cai *et al.* assessed complications of sickle cell disease by detecting sea fan neovascularization in funduscopic images, which is a vision-threatening complication warranting prophylactic management.[96] Doan *et al.* evaluated therapy response in ALL patients by using DL methods to detect residual lymphoblasts after receiving induction chemotherapy.[97] Finally, DL models for relapse prediction using baseline imaging have been developed for extranodal natural killer/T-cell lymphoma[98] and mantle cell lymphoma.[99] However, further evaluations upon external cohorts are needed for these advanced stage tasks.

Conclusions

The use of deep learning in hematologic conditions has attracted significant interest in recent years. As noted above, researchers have utilized multiple data structures including radiologic images, pathology specimens, clinical data, and atypical imaging such as funduscopic examinations to perform a variety of clinically relevant tasks. Most Authors reported high model performance for disease diagnosis, segmentation, and subtyping. Other studies explored tasks beyond human capabilities such as genomic inference and prognostication from imaging analysis alone. Few studies have used hematologic conditions as a means to implement state-of-the-art architectures to improve the field of DL in general. Compared to other clinical domains, DL in hematology is still in its infancy, so it is not widely used in clinical practice. As such, the intention of this review is to introduce broad concepts to hematology clinicians to assist in the evaluation and understanding of future DL implementations, as well as to provide an overview of the clinical uses currently being explored throughout patient care.

The fact that it is still early days for DL in hematology may be due to a lack of appropriate algorithm design, data availability, computational resources, and insufficient disease-specific expertise involved in DL development.[100] To the best of our knowledge, there are still no large clinically-annotated multi-modal public datasets for many hematologic conditions. In addition, critical morphological information in hematopathology may only be available at higher magnification levels, surpassing the limits of standard pathology scanners. Although these structural barriers continue to compromise the development of DL in hematology, rapid technological advances continue, and interest for DL within the academic community is growing.[101]

Though promising, the methods and conclusions from the numerous studies are heterogenous and challenging to compare. As yet, there is no standardized approach in DL research, reporting, or implementation. In the present overview, the majority of publications were evaluated by internal validation strategies, with the minority evaluated on external institution cohorts. Explaining model predictions were not ubiquitous, and few DL models were compared directly against human evaluation. Major government initiatives currently aim to standardize DL protocol design,[102] and, despite the variance in outcome reporting in DL analyses, the SPIRIT-AI, STARD-AI, and CONSORT-AI initiatives aim to standardize future clinical trial design and reporting of artificial intelligence interventions.[103-105]

The research and results of DL analyses must be interpreted cautiously, as a number of practical and ethical issues have arisen in other domains of machine learning. CNN are prone to “memorize” the training set; thus, the initial high performance may fail to be carried forward on new previously unseen data. For this reason, it is imperative to evaluate DL models on external cohorts from separate institutions. If training data are acquired from multiple institutions, care must be given to correct for known “batch effects,” as DL models may infer site-specific artifact signatures not related to the underlying disease biology.[106] Similarly, researchers should investigate explainability and error analysis to ensure that the models rely on scientifically reasonable features and ignore irrelevant factors. In addition, uncertainty in model predictions are rarely reported but are arguably necessary for clinical implementation of DL algorithms.

In this review, the majority of DL applications are aimed towards earlier phases of clinical care, such as automation and disease detection. DL in lymphoma resulted in the plurality of exploratory analyses, likely due to the importance of both radiologic and pathologic findings in the care of lymphoma patients. Though explored in a myriad of malignant and non-malignant conditions, notably lacking are DL applications in stem cell transplantation and many other non-malignant processes where morphological assessment is paramount, such as thrombotic microangiopathies.

Future work is needed to address large scale applications of DL in hematology. As a hematopathologist typically assesses histology specimens at different magnification levels, customized architectures to implement multi-scale image analysis should be explored. DL in solid oncology is widely used, in part due to the publicly available digital biopsy specimens provided by The Cancer Genome Atlas,[107] of which there is no analogous database for hematologic conditions. In addition, the combination of multi-modal data structures that incorporate images in concert with flow cytometry, molecular analyses, cytogenetics, or other clinical factors may provide additional relevant features to improve DL models.

While numerous considerations remain before large-scale implementation of DL is feasible, the development of new models and applications in hematology is rapidly increasing, and it is imperative for clinicians to be aware of the opportunities that DL may provide.

Disclosures

ATP reports support via grants from NIH/NCI U01-CA243075, grants from NIH/NIDCR R56-DE030958, grants from Horizon 2021-SC1-BHC, grants from DoD Breakthrough Cancer Research program BC211095, and grants from SU2C (Stand Up to Cancer) Fanconi Anemia Research Fund – Farrah Fawcett Foundation Head and Neck Cancer Research Team Grant during the conduct of the study. ATP reports grants from Abbvie via the UChicago – Abbvie Joint Steering Committee Grant, and grants from Kura Oncology. ATP reports personal fees from Prelude Therapeutics Advisory Board, from Elevar Advisory Board, from Ayala Advisory Board, personal fees from Abbvie Advisory Board, and from Privo Advisory Board outside of submitted work. MES is an employee of Sonic Healthcare. AS has no conflicts of interest to disclose.

Contributions

All authors conceived the manuscript. AS wrote the manuscript. All authors approved the final version.

Funding

ATP reports this study has been funded in whole or in part with Federal funding by the NCI-DOE Collaboration established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health, Cancer Moonshot Task Order N. 75N91019F00134 and under Frederick National Laboratory for Cancer Research Contract 75N91019D00024. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357.

Data-sharing statement

No applicable.

References

1. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301-1309.
2. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054-1056.
3. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559-1567.
4. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1(8):789-799.
5. Saillard C, Schmauch B, Laifa O, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology. 2020;72(6):2000-2013.
6. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices#resources. Accessed November 3, 2021.
7. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. An expert-annotated dataset of bone marrow cytology in hematologic malignancies. The Cancer Imaging Archive; 2021.
8. Matek C, Schwarz S, Marr C, Spiekermann K. A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls. The Cancer Imaging Archive; 2019.
9. Schouten JPE, Matek C, Jacobs LFP, Buck MC, Bosnacki D, Marr C. Tens of images can suffice to train neural networks for malignant leukocyte detection. Sci Rep. 2021;11(1):7995.
10. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nature Machine Intelligence. 2019;1(11):538-544.
11. Matek C, Krappe S, Munzenmayer C, Haferlach T, Marr C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. Blood. 2021;138(20):1917-1927.
12. Sidhom JW, Siddarthan IJ, Lai BS, et al. Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. NPJ Precis Oncol. 2021;5(1):38.
13. Eckardt JN, Schmittmann T, Riechert S, et al. Deep learning identifies acute promyelocytic leukemia in bone marrow smears. BMC Cancer. 2022;22(1):201.
14. Eckardt JN, Middeke JM, Riechert S, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. Leukemia. 2022;36(1):111-118.
15. Chollet F. Deep learning with Python. 1st ed. USA: Manning Publications Co.; 2017.
16. Murphy KP. Probabilistic machine learning: an introduction. The MIT Press; 2022.

17. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: Users' Guides to the medical literature. JAMA. 2019;322(18):1806-1816.
18. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7(1):16878.
19. Dolezal J, Kochanny S, Howard F. Slideflow: a unified deep learning pipeline for digital histology: Zenodo; 2022.
20. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7-12 2015; Boston, MA, USA. IEEE; c2015. p. 1-9.
21. Olah C, Satyanarayan A, Johnson I, et al. The building blocks of interpretability. Distill. 2018;3(3).
22. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 20-25 June 2009; Miami, FL, USA. IEEE; c2009. p. 248-255.
23. Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. Med Image Anal. 2021;70:102032.
24. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. 3 Jun 2021. doi: 10.48550/arXiv.2010.11929 [preprint, not peer reviewed].
25. Bianco S, Cadene R, Celona L, Napoletano P. Benchmark analysis of representative deep neural network architectures. IEEE access. 2018;6:64270-64277.
26. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504-507.
27. Combalia M, Vilaplana V. Monte-Carlo sampling applied to multiple instance learning for histological image classification. In: Stoyanov D, Taylor Z, Carneiro G, et al., eds. Deep learning in medical image analysis and multimodal learning for clinical decision support; 20 Sep 2018; Granada, Spain. Springer International Publishing; c2018. p. 274-281.
28. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence. 1997;89(1-2):31-71.
29. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning; 10-15 July 2018; Stockholm, Sweden. PMLR; c2018. p. 2127-2136.
30. Sadafi A, Makhro A, Bogdanova A, et al. Attention based multiple instance learning for classification of blood cell disorders. In: Medical image computing and computer assisted intervention – MICCAI 2020; 4-8 Oct 2020; Lima, Peru. Springer International Publishing; c2020. p. 246-256.
31. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555-570.
32. Hooker S, Erhan D, Kindermans P-J, Kim B. A benchmark for interpretability methods in deep neural networks. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); Vancouver, Canada. c2019. p. 9734-9745.
33. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? In: Advances in neural information processing systems; c2021. p. 12116-12128.
34. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750.
35. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv. 2017 Jun 12. doi: 10.48550/arXiv.1706.03825. [preprint, not peer reviewed].
36. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215.
37. Krause J, Grabsch HI, Kloor M, et al. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. J Pathol. 2021;254(1):70-79.
38. Fan H, Zhang F, Xi L, Li Z, Liu G, Xu Y. LeukocyteMask: an automated localization and segmentation method for leukocyte in blood smear images using deep neural networks. J Biophotonics. 2019;12(7):e201800488.
39. Alam MM, Islam MT. Machine learning approach of automatic identification and counting of blood cells. Healthc Technol Lett. 2019;6(4):103-108.
40. Vajen B, Hanselmann S, Lutterloh F, et al. Classification of fluorescent R-Band metaphase chromosomes using a convolutional neural network is precise and fast in generating karyograms of hematologic neoplastic cells. Cancer Genet. 2022;260-261:23-29.
41. Jemaa S, Fredrickson J, Carano RAD, Nielsen T, de Crespigny A, Bengtsson T. Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. J Digit Imaging. 2020;33(4):888-894.
42. Xu L, Tetteh G, Lipkova J, et al. Automated whole-body bone lesion detection for multiple myeloma on (68)Ga-Pentixafor PET/CT imaging using deep learning methods. Contrast Media Mol Imaging. 2018;2018:2391925.
43. Weisman AJ, Kieler MW, Perlman SB, et al. Convolutional neural networks for automated PET/CT detection of diseased lymph node burden in patients with lymphoma. Radiol Artif Intell. 2020;2(5):e200016.
44. Weisman AJ, Kim J, Lee I, et al. Automated quantification of baseline imaging PET metrics on FDG PET/CT images of pediatric Hodgkin lymphoma patients. EJNMMI Phys. 2020;7(1):76.
45. Sadik M, Lopez-Urdaneta J, Ulen J, et al. Artificial intelligence could alert for focal skeleton/bone marrow uptake in Hodgkin's lymphoma patients staged with FDG-PET/CT. Sci Rep. 2021;11(1):10382.
46. Yang Y, Tang Y, Gao R, et al. Validation and estimation of spleen volume via computer-assisted segmentation on clinically acquired CT scans. J Med Imaging (Bellingham). 2021;8(1):014004.
47. Liu W, Liu M, Guo X, et al. Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning. Eur Radiol. 2020;30(6):3567-3575.
48. de Haan K, Ceylan Koydemir H, Rivenson Y, et al. Automated screening of sickle cells using a smartphone-based microscope and deep learning. NPJ Digit Med. 2020;3(1):76.
49. Shaw M, Claveau R, Manescu P, et al. Optical mesoscopy, machine learning, and computational microscopy enable high information content diagnostic imaging of blood films. J Pathol. 2021;255(1):62-71.
50. Huber N, Anderson T, Missert A, et al. Clinical evaluation of a phantom-based deep convolutional neural network for whole-body-low-dose and ultra-low-dose CT skeletal surveys. Skeletal Radiol. 2022;51(1):145-151.
51. Theruvath AJ, Siedek F, Yerneni K, et al. Validation of deep learning-based augmentation for reduced (18)F-FDG dose for PET/MRI in children and young adults with lymphoma. Radiol Artif Intell. 2021;3(6):e200232.
52. Rajaraman S, Antani SK, Poostchi M, et al. Pre-trained

convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018;6:e4568.

53. Rajaraman S, Silamut K, Hossain MA, et al. Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. J Med Imaging (Bellingham). 2018;5(3):034501.
54. Kuo PC, Cheng HY, Chen PF, et al. Assessment of expert-level automated detection of plasmodium falciparum in digitized thin blood smear images. JAMA Netw Open. 2020;3(2):e200206.
55. Manescu P, Shaw MJ, Elmi M, et al. Expert-level automated malaria diagnosis on routine blood films with deep neural networks. Am J Hematol. 2020;95(8):883-891.
56. Li S, Du Z, Meng X, Zhang Y. Multi-stage malaria parasite recognition by deep learning. Gigascience. 2021;10(6):giab040.
57. Lee SY, Chen CME, Lim EYP, et al. Image analysis using machine learning for automated detection of hemoglobin H inclusions in blood smears - a method for morphologic detection of rare cells. J Pathol Inform. 2021;12:18.
58. Doan M, Sebastian JA, Caicedo JC, et al. Objective assessment of stored blood quality by deep learning. Proc Natl Acad Sci U S A. 2020;117(35):21381-21390.
59. Acevedo A, Merino A, Boldu L, Molina A, Alferez S, Rodellar J. A new convolutional neural network predictive model for the automatic recognition of hypogranulated neutrophils in myelodysplastic syndromes. Comput Biol Med. 2021;134:104479.
60. Kimura K, Ai T, Horiuchi Y, et al. Automated diagnostic support system with deep learning algorithms for distinction of Philadelphia chromosome-negative myeloproliferative neoplasms using peripheral blood specimen. Sci Rep. 2021;11(1):3367.
61. Sahlol AT, Kollmannsberger P, Ewees AA. Efficient classification of white blood cell leukemia with improved swarm optimization of deep features. Sci Rep. 2020;10(1):2536.
62. Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technol Cancer Res Treat. 2018;17:1533033818802789.
63. Mori J, Kaji S, Kawai H, et al. Assessment of dysplasia in bone marrow smear with convolutional neural network. Sci Rep. 2020;10(1):14734.
64. Sirinukunwattana K, Aberdeen A, Theissen H, et al. Artificial intelligence-based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in MPN patients. Blood Adv. 2020;4(14):3284-3294.
65. Gehlot S, Gupta A, Gupta R. A CNN-based unified framework utilizing projection loss in unison with label noise handling for multiple myeloma cancer diagnosis. Med Image Anal. 2021;72:102099.
66. Li D, Bledsoe JR, Zeng Y, et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nat Commun. 2020;11(1):6004.
67. Syrykh C, Abreu A, Amara N, et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. NPJ Digit Med. 2020;3(1):63.
68. Sibille L, Seifert R, Avramovic N, et al. (18)F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. Radiology. 2020;294(2):445-452.
69. Zhou Z, Jain P, Lu Y, et al. Computer-aided detection of mantle cell lymphoma on (18)F-FDG PET/CT using a deep learning convolutional neural network. Am J Nucl Med Mol Imaging. 2021;11(4):260-270.
70. Huang SC, Kothari T, Banerjee I, et al. PENet-a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ Digit Med. 2020;3(1):61.
71. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep. 2020;10(1):22147.
72. Kainz B, Heinrich MP, Makropoulos A, et al. Non-invasive diagnosis of deep vein thrombosis from ultrasound imaging with machine learning. NPJ Digit Med. 2021;4(1):137.
73. Kwon JM, Cho Y, Jeon KH, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Health. 2020;2(7):e358-e367.
74. Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng. 2020;4(1):18-27.
75. Li S, Yang Q, Jiang H, Cortes-Vecino JA, Zhang Y. Parasitologist-level classification of apicomplexan parasites and host cell with deep cycle transfer learning (DCTL). Bioinformatics. 2020;36(16):4498-4505.
76. Kimura K, Tabe Y, Ai T, et al. A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. Sci Rep. 2019;9(1):13385.
77. Ahmed N, Yigit A, Isik Z, Alpkocak A. Identification of leukemia subtypes from microscopic images using convolutional neural network. Diagnostics (Basel). 2019;9(3):104.
78. Huang F, Guang P, Li F, Liu X, Zhang W, Huang W. AML, ALL, and CML classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network: a STARD compliant diagnosis research. Medicine (Baltimore). 2020;99(45):e23154.
79. Achi HE, Belousova T, Chen L, et al. Automated diagnosis of lymphoma with digital pathology images using deep learning. Ann Clin Lab Sci. 2019;49(2):153-160.
80. Guan Q, Wan X, Lu H, et al. Deep convolutional neural network Inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study. Ann Transl Med. 2019;7(14):307.
81. Mohlman JS, Leventhal SD, Hansen T, Kohan J, Pascucci V, Salama ME. Improving augmented human intelligence to distinguish Burkitt lymphoma from diffuse large B-cell lymphoma cases. Am J Clin Pathol. 2020;153(6):743-759.
82. Miyoshi H, Sato K, Kabeya Y, et al. Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma. Lab Invest. 2020;100(10):1300-1310.
83. Yun J, Park JE, Lee H, Ham S, Kim N, Kim HS. Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma. Sci Rep. 2019;9(1):5746.
84. Zhao J, Zhang M, Zhou Z, Chu J, Cao F. Automatic detection and classification of leukocytes using convolutional neural networks. Med Biol Eng Comput. 2017;55(8):1287-1301.
85. Lippeveld M, Knill C, Ladlow E, et al. Classification of human white blood cells using machine learning for stain-free imaging flow cytometry. Cytometry A. 2020;97(3):308-319.
86. Wu YY, Huang TC, Ye RH, et al. A hematologist-level deep learning algorithm (BMSNet) for assessing the morphologies of single nuclear balls in bone marrow smears: algorithm development. JMIR Med Inform. 2020;8(4):e15963.
87. Durant TJS, Olson EM, Schulz WL, Torres R. Very deep convolutional neural networks for morphologic classification of erythrocytes. Clin Chem. 2017;63(12):1847-1855.
88. Xu M, Papageorgiou DP, Abidi SZ, Dao M, Zhao H, Karniadakis GE. A deep convolutional neural network for classification of red blood cells in sickle cell anemia. PLoS Comput Biol. 2017;13(10):e1005746.
89. Zhou Y, Yasumoto A, Lei C, et al. Intelligent classification of

platelet aggregates by agonist type. Elife. 2020;9:e52779.

90. Rehman A, Abbas N, Saba T, Rahman SIU, Mehmood Z, Kolivand H. Classification of acute lymphoblastic leukemia using deep learning. Microsc Res Tech. 2018;81(11):1310-1317.
91. Bruck OE, Lallukka-Bruck SE, Hohtari HR, et al. Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with MDS. Blood Cancer Discov. 2021;2(3):238-249.
92. Swiderska-Chadaj Z, Hebeda KM, van den Brand M, Litjens G. Artificial intelligence to detect MYC translocation in slides of diffuse large B-cell lymphoma. Virchows Arch. 2021;479(3):617-621.
93. Irshaid L, Bleiberg J, Weinberger E, et al. Histopathologic and machine deep learning criteria to predict lymphoma transformation in bone marrow biopsies. Arch Pathol Lab Med. 2022;146(2):182-193.
94. Jullien M, Tessoulin B, Ghesquieres H, et al. Deep-learning assessed muscular hypodensity independently predicts mortality in DLBCL patients younger than 60 years. Cancers (Basel). 2021;13(18):4503.
95. Cahan N, Marom EM, Soffer S, et al. Weakly supervised attention model for RV strain classification from volumetric CTPA scans. Comput Methods Programs Biomed. 2022;220:106815.
96. Cai S, Parker F, Urias MG, Goldberg MF, Hager GD, Scott AW. Deep learning detection of sea fan neovascularization from ultra-widefield color fundus photographs of patients with sickle cell hemoglobinopathy. JAMA Ophthalmol. 2021;139(2):206-213.
97. Doan M, Case M, Masic D, et al. Label-free leukemia monitoring by computer vision. Cytometry A. 2020;97(4):407-414.
98. Guo R, Hu X, Song H, et al. Weakly supervised deep learning for determining the prognostic value of (18)F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type. Eur J Nucl Med Mol Imaging. 2021;48(10):3151-3161.
99. Lisson CS, Lisson CG, Mezger MF, et al. Deep neural networks and machine learning radiomics modelling for prediction of relapse in mantle cell lymphoma. Cancers (Basel). 2022;14(8):2008.
100. Kochanny SE, Pearson AT. Academics as leaders in the cancer artificial intelligence revolution. Cancer. 2021;127(5):664-671.
101. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. Lancet Haematol. 2020;7(7):e541-e550.
102. New NCI-DOE collaboration project, IMPROVE, seeks deep learning model approaches. https://datascience.cancer.gov/news-events/news/new-nci-doe-collaboration-project-improve-seeks-deep-learning-model-approaches. Accessed May 21, 2022.
103. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-1363.
104. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374.
105. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. 2021;11(6):e047709.
106. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021;12(1):4423.
107. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-1120.