WHITE PAPER

iMerit

THE CHALLENGE OF MED AI ANNOTATION

DEMAND FOR DEPLOYMENT

Although a year ago may now seem like the distant past in the tech world, the promise of artificial intelligence and machine learning has not abated. In healthcare, while advances continue to be largely in the research and development, medical practitioners are increasingly looking to AI to play a greater role in healthcare than they did even a year ago. One of the questions uppermost on the minds of clinician attendees at radiologic conferences in 2019 was, how long before AI takes my job? While some of those worries may have faded rapidly, in part due to the coronavirus pandemic which has demonstrated how unprepared most world health systems are for a large-scale emergency, the demand for deployment of AI across a host of medical disciplines has only grown. Can AI bring healthcare the coveted triple win proposition, can it be good for providers, patients, and payers? Many believe it can.

Unlike other market segments—transportation, entertainment, manufacturing, and commerce, for example—accelerating AI deployments in healthcare cannot be accomplished simply by throwing more resources at the problem. AI models require vast amounts of data to learn to navigate in the real world. Whether the data is in the form of text or images, it must often first be annotated into a language the algorithm can understand. This work is often performed by humans—domain experts—who are trained to label relevant data elements in literally millions of images. Accuracy is of utmost importance, for the old computer adage “garbage in, garbage out” applies even more strongly to AI than it does to conventional computer programming. Correcting a faulty AI model is not simply a matter of rewriting a few lines of code. The entire learning process may have to be restarted from scratch.

Where do domain experts come from? The words “expert” and “experience” share the same root, and indeed, to become a domain expert one must acquire an intimate familiarity with the particular segment of the world the model requires. For autonomous vehicles, that expertise is widespread. Virtually all drivers (and even many non-drivers) have the necessary level of familiarity to annotate images relevant to this domain. We begin cultivating that expertise as children, looking out the car window as our parents drive us around town. The ubiquity of expertise, the relative ease of recruiting domain experts, and hence the resultant high volume of training data able to be generated is an important reason why autonomous vehicles are making such rapid progress among real world applications of AI technology.

Computer Vision use cases in Medical AI

Domain	Medium	Example use cases
Physical exam	Static images	■ Dermatologic lesion identification ■ Retinopathy identification in ophthalmology
Endoscopy, minimally invasive surgery, & robotic surgery	Video	■ Pathology identification in esophagogastroduodenoscopy
Digital radiology	Static images, video, 3D & 4D	■ Cancer detection in CTs ■ Gallbladder inflammation on ultrasound
Digital pathology	Static images	■ Abnormal cells on whole slide imaging
Sports medicine / Biomechanics / Behavioral medicine	Static images, video	■ Injury prevention in athletes ■ Fall prevention in nursing homes

Source: **iMerit**

From the preceding example it is easy to see why medical AI faces a critical shortage of domain experts, and will continue to face a shortage in the future. Unlike domains in the external world, medical domains are hidden from view. Even medical students attain their expertise over years of study, and continue to hone their skills long after they gain certification. Whether looking at radiologic images, pathologic slides, or surgical videos, full depth of understanding resides in the hands of a relative few. Relying solely on medical professionals or students to annotate images of tumors in a CT scan, for example, is not a scalable strategy to meet the demand that is already here, much less future demand. These individuals are limited in number and have limited time to devote to image annotation, leading to lengthy project timelines and high costs.

But what if it were possible to train lay people to annotate medical images? The strategy is not as radical as it might seem. Consider the field of civil engineering. While it takes years of training to design a bridge, bridge inspectors can quickly identify weak rivets, cracked footings, rusted beams, and other dangerous signs of wear, all without the advanced training required of certified engineers.

Domains involving medical imaging, such as radiology, pathology, endoscopy, or robotic surgery, can be far more heterogeneous than civil engineering, which helps explain why engineering technology like computer aided design (CAD) is ubiquitous, while computer assistance still plays only a marginal role in clinical practice.

Cognitive load requirements for radiological use cases

Approaching High SME Domains in Digital Radiology first involves identifying key features in the cognitive load by modality. These features include:

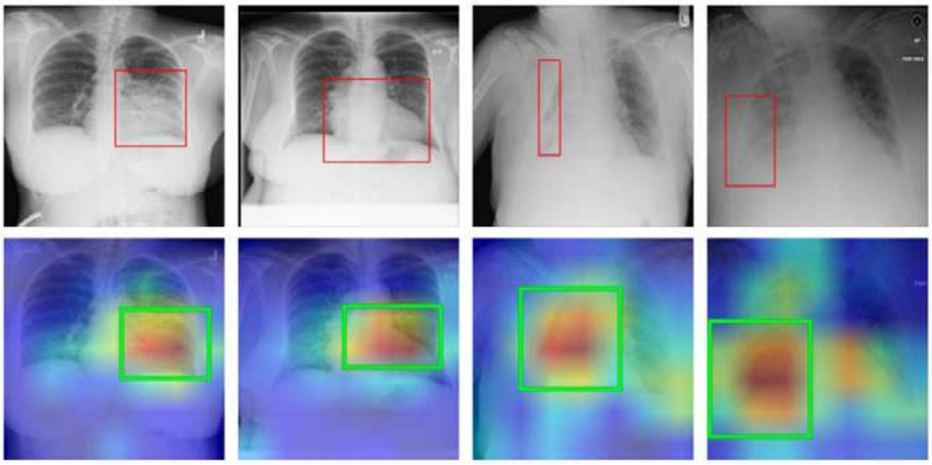
- Anatomy / Physiology: Breadth and complexity of relevant organ systems & structures
- Pathology / Pathophysiology: Frequency, variability, and complexity of abnormalities
- Visualization, Navigation, & Spatial Reasoning: Degree of difficulty in manipulating imaging including Signal versus Noise
- Ontologic Complexity: Number of factors and complexity of relationships in a decision tree.

In this chart, we have evaluated the cognitive load required to complete these tasks 1 to 7, 1 being the lowest cognitive load required.

Modality	Brief description	Used to diagnose	Cognitive Load			
			Anatomy/ Physiology	Pathology/ Pathophysiology	Visualization, Navigation, & Spatial Reasoning	Ontologic Complexity
X-ray	Ionizing radiation projected through to capture structures on static film.	■ Bone fractures ■ Arthritis ■ Osteoporosis ■ Breast cancer ■ Swallowed items	1	1	2	1
CT Scan	Stacked X-ray in an array to create a three dimensional set of images.	■ Injuries from trauma ■ Tumors and cancers ■ Vascular disease ■ Heart disease ■ Infections	4	4	3	3
MRI	Magnetic Fields used to visualize soft tissue structures. Often utilizing multiple protocols and potentially at multiple time points.	■ Aneurysms ■ Multiple sclerosis (MS) ■ Stroke ■ Spinal cord disorders ■ Blood vessel issues	7	7	6	6
Ultrasound	Sound waves used to create videos of 2 dimensional cross-sections of tissue.	■ Gallbladder disease ■ Breast lumps ■ Genital/prostate issues ■ Blood flow problems ■ Monitoring pregnancy	3	3	5	1

Source: iMerit

Expert vs. AI

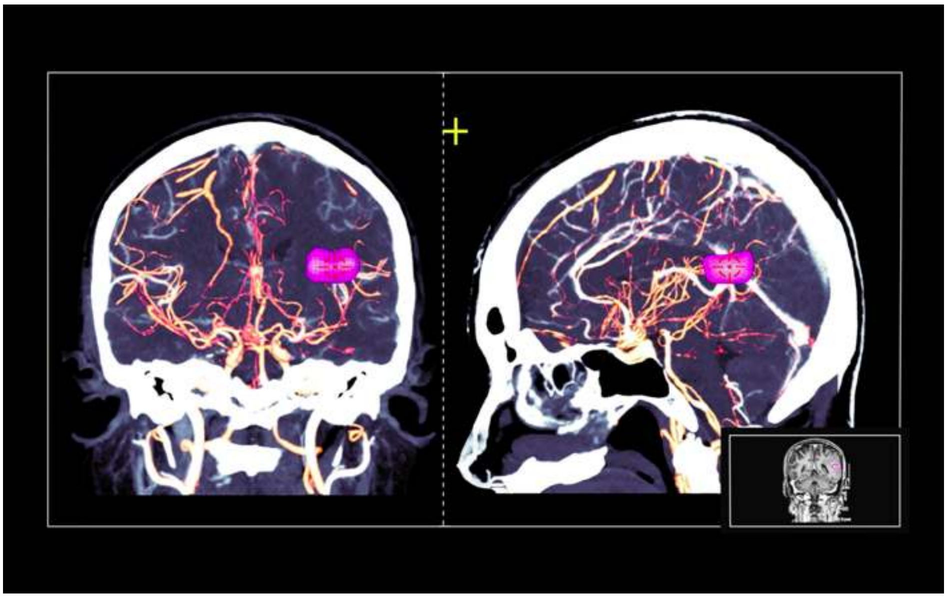


Source: **Diagnose Like a Radiologist**

The top row shows four chest X-rays with lesions annotated in bounding boxes by human experts. The bottom row shows heat maps generated by an algorithm through a process called mask inference. Higher response (i.e., greater presence of lesions) are in red. These heat maps may then be fed back into the original images to generate a more localized and clearer image of the lesion. The expert annotated images would not be used in this part of the process. Eventually the expert annotated ground truth would not be needed.

Medical AI does not suffer from a lack of data. Millions upon millions of images—of internal organs, tumors, fetuses, skeletal components, dental features, retinal scans, etc.—reside in data banks in every city, state, and country. In addition, the increased use of video-driven technology for robotic, procedural, and non-invasive surgery adds to the "digitization" of clinical imaging. What secrets do these images hold? What similarities and differences might prompt researchers to devise new treatments, or clinicians to diagnose serious illnesses before they caused irreparable harm or death to the patient? Only time will tell.

Data by itself is neither information, nor knowledge. This view can be counter-intuitive in a clinical setting. An oncologist might insist that a CT scan, for example, contains information which is absolutely vital. But this is true only because the radiologist brings a vast amount of experience to the viewing of the image. Without that experience the image is not useful. "Data is a set of discrete, objective facts about events," writes data scientist Thomas Davenport. For data to be useful, he writes, it requires context, categorization, and other attributes only human input can provide. This is the role of the annotator, and it is critical. The job is made more challenging by the fact that even the "objectivity" of a CT scan or other imaging data can be called into question. Just as different computer monitors vary in color and display differences from subtle to extreme, different scanning systems may assign different gray-scale values to similar structures, leading to edge cases calling for expert evaluation. Does this gray area represent healthy tissue or diseased tissue? These questions must be answered before an image can be rendered mathematically to train a model. As in any endeavor, perfection is impossible to achieve, but greater accuracy in the annotative process results in far more efficiency in the iterative machine learning procedure.



Source: **iMerit**

Annotation of cerebral vasculature requires understanding of topology in multiple planes. Above, CT Angiography is evaluated in coronal and sagittal 3D projections with lesions annotated in single cross-sections.

One of the buzzwords current today is “explainable” AI. Neural networks do their work not exactly in secret, but in ways that are definitely mysterious. If an AI-powered system flags a CT scan or X-ray as problematic, while the radiologist does not see a problem, it is a reasonable question to ask, what’s going on here? Sometimes, however, the algorithm simply cannot give up its reasoning, largely because it does not reason as human beings do. AI machines, while very good at analyzing images and identifying similarities and differences, are generally not good at identifying cause and effect. In addition, because of the relative scarcity of well-annotated medical images, there is widespread utilization of so-called “weak supervision” in the data being used to train medical AI models. This leads to a further mistrust of AI diagnoses among clinicians. In fact, while AI is generally seen as lightening the workload and improving the efficiency of radiologists, the lack of explainable AI has in some cases done exactly the opposite. When the algorithm and radiologist disagree, the doctor may devote substantial time and effort not to treating the patient, but to trying to understand why the algorithm has come to its conclusion.

CONTEXT IS KING

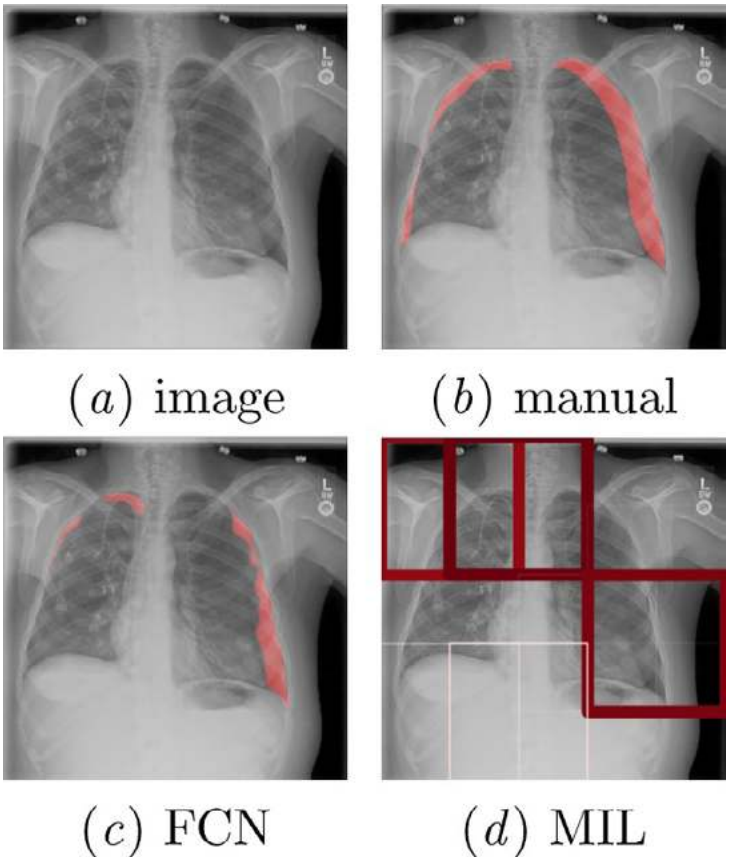
A little over 100 years ago the philosopher Ludwig Wittgenstein very presciently suggested a thought experiment which presaged one of the ways computer vision has been designed to work. Wittgenstein imagined “a white surface with irregular black spots on it.” He then imagined that surface covered with a fine square mesh, “and then saying of every square whether it is black or white.” This method allows him to completely describe the picture, but, he warns, “tells us nothing about the picture.” Wittgenstein compares the mesh to Newtonian mechanics, for example, which can be used to describe the world – but only according to Newtonian mechanics. So while Newtonian mechanics is perfectly able to describe the actions of billiard balls on a table, it cannot tell us that these rolling, colliding objects are part of a game called “billiards”.

The act of overlaying an image with a mesh, and assigning each square a value, is perfectly analogous to the annotated images used to train AI models. The model gives us a description of the image, but what is really needed by the clinician is context. Unless the clinician can trust the AI to be more than simply a descriptive overlay, the technology will prove to be of limited use in medical practice. Locally annotated training data, such as that performed by human experts, can provide the necessary context for the algorithm to deliver trustworthy results. However, images for models with weak supervision (lacking ground truth) are more widely available through open data sources on the internet. Researchers are experimenting with various methods of using deep learning to train algorithms without or with limited ground truth annotations. These efforts often require large volumes of data, however, the obtaining of which can lead to regulatory and data siloing barriers.

Researchers generally agree there are three levels of image data annotation: classification level, bounding box level, and pixel level. In medical applications, the classification level identifies areas of the body such as a chest X-ray, endoscope, etc. The bounding box level highlights a general area for diagnostic purposes in a rectangle, such as a darkened area on the lung or erosion on an EGD. The pixel level drills down to outline the area and enable measurement and segmentation of the diseased tissue. The challenge is to train the machine to identify and annotate ground truths from classification level images alone.

One method that has shown promise is utilizing Multi-Instance Learning (MIL). This method, developed over a decade ago as a computer vision training technique that has been widely used in entertainment and gaming applications, involves breaking up an image into a set of parts (“instances”), which are called image patches. In the initial training stage, locally annotated images are used. In training an algorithm to identify a pneumothorax, for example, each patch is given a value of positive or negative (1 or 0) depending on whether it showed pneumothorax or not. The machine is able to learn from the ground truth data, but the goal is to provide only weakly annotated data to the algorithm and have it not only identify pneumothorax, but also provide a basis for its identification—explainability. This is achieved through additional MIL on the patches, essentially breaking them up into smaller patches for deeper analysis. This further analysis produces a “heat map”—a visual representation of the features the algorithm identified which prompted its diagnosis of pneumothorax. In this way, the clinician has not just a “black box” result, but a road map back through the iterative process which led to the diagnosis. MIL can be used to identify multiple conditions as well.

Multi-instance Learning



Source: **Pneumothorax Detection and Localization in Chest Radiographs**

(a) Original image (b) Manual expert labeling (c) Fully Convolutional Neural Network (AI-generated)
(d) Multi-instance Learning with the “image patches” visible

While this research is still in preliminary stages, the current work is encouraging. Downsides include a lack of clarity in some of the heat maps and other imaging created from the patches. But the promise of increased explainability is there. Future developments may combine these MIL advances with natural language processing (NLP) applications to automatically generate preliminary reports for clinicians. These reports would include the "reasoning" behind the algorithm's conclusions, though it is important to understand that AI explainability is based on ever deeper analysis of visual data, not human style deduction. Nonetheless if it is sufficient to provide clinicians with confidence in relying more on AI, it would go a long way to ensure better patient outcomes and help clinicians better utilize their time and expertise.

While most of the headlines involving AI and COVID-19 concern the search for a vaccine or effective drugs, medical imaging algorithms have also played an important role in the fight against the virus.

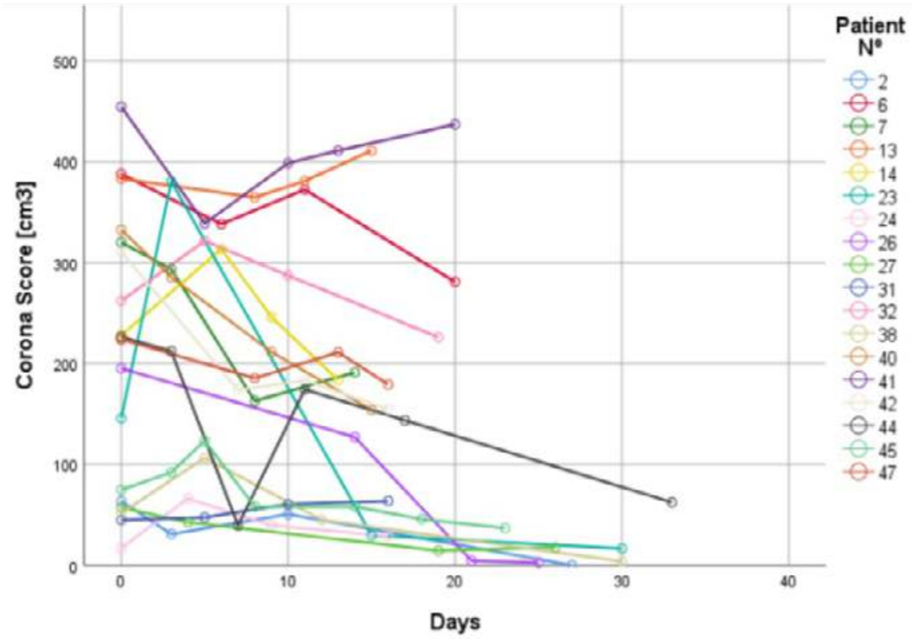
Explainability is important even when the disease under examination is itself more novel. This is the case with COVID-19, a mystery wrapped in an enigma wrapped in a virus. Among the (as yet) unexplainable phenomena surrounding the novel coronavirus are such important questions as, why do the majority of infected individuals show minor or no symptoms, while others become gravely or mortally ill? Why do some people seemingly near recovery suffer a sudden and severe relapse? What are the connections between the coronavirus and liver, kidney and blood disorders? And can AI help fight this illness given the paucity and heterogeneity of ground truth data?

While most of the headlines involving AI and COVID-19 concern the search for a vaccine or effective drugs, medical imaging algorithms have also played an important role in the fight against the virus. An international team from Israel, China, and the United States has had success in training an algorithm to detect "ground-glass nodules," a distinctive feature in the lungs of patients suffering from the coronavirus, so named because of their appearance on chest CT images . The researchers compare their new algorithm, based on existing algorithms used in the detection and treatment of lung disease, to the novel coronavirus itself which is presumed to have mutated from earlier forms. And in China, CT scans were used for screening of symptomatic patients and showed higher sensitivity than molecular testing using RT-PCR.

The challenges presented by novel coronavirus include limited data, limited time—since rapid deployment is essential if the algorithm is to be useful in controlling the spread of the disease—and the need for a global solution. As we have seen, the virus does not respect borders or populations.

Going back over existing images from China that predated the pandemic, and which were used to train the original algorithm, the team discovered some instances of ground-glass nodules, which they were able to confirm were in fact COVID-19 related. Using innovative computer imaging techniques which included "slicing" the 3D images into 2D scans for faster results, they were able to train the new algorithm to detect the ground glass nodules in new cases. Using additional data from these patients they generated a "Corona score" to quantify the progress of the disease.

AI-generated 'Corona Score'



Source: **Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic**

The AI-generated 'Corona Score' based on a variety of data, including imaging and other clinical data. The higher the score, the more severe the symptoms of COVID-19. The graph clearly shows marked differences in the disease's progress among patients.

The team proposes to expand the capabilities of this initial algorithm to detect signs of COVID-19 in additional affected areas beside the lungs. One of the most significant aspects of this research, according to the team, is that the AI models and the radiologists are learning about this serious condition in tandem. They see a future AI contribution to the treatment of COVID-19 in three areas: Detection, enabling automatic triage capabilities to flag suspected positive cases; Measurement and the ability to quantify findings; and the Tracking of Disease Progression, enabling decision support for patient management. While these are all areas clinicians are currently managing without AI, the rapid spread of COVID-19 and the subsequent stress placed upon healthcare facilities and personnel argue strongly in favor of AI in the clinic, provided it can deliver accurate results.

But even with image augmentation and other advanced techniques, expert-annotated ground truth data will remain the bedrock of AI imaging for the foreseeable future. But if the supply of experts is limited to trained medical personnel, the future may not be a sustainable one. The key to training an effective workforce of lay domain specialists in the medical imaging field is multifaceted. Even before a training curriculum can be designed, a methodology must be devised to identify and evaluate prospective candidates who demonstrate both the aptitude and the motivation to complete a rigorous training regime. Then experienced medical professionals must be engaged to design the training curriculum, teaching both the general skills of using the annotation tools, and the specific skills associated with the project focus. Ideally the curriculum would be designed in close consultation with the client, resulting in a highly focused, narrow but deep course of study. And it is absolutely critical that a feedback component be in place to review the work of the annotators and provide higher level guidance on edge cases—the "human-in-the-loop." We will examine these requirements individually.

IDENTIFYING APTITUDE

Developing expertise in annotating medical images relies on very different skill sets than the ones needed for expertise in a clinical setting. First among these is pattern recognition. Some critics of medical AI minimize the importance of this skill, going so far as to charge that the "intelligence" of neural networks is no more than pattern recognition honed through iteration to a high degree of accuracy. Whether a machine's ability to identify a cancerous tumor is indicative of "intelligence" or of something else does not devalue the role the machine can play in improving patient care. While an algorithm can learn to recognize similarities and differences among thousands of images, it is dependent on human input to do so with a reliable degree of accuracy.

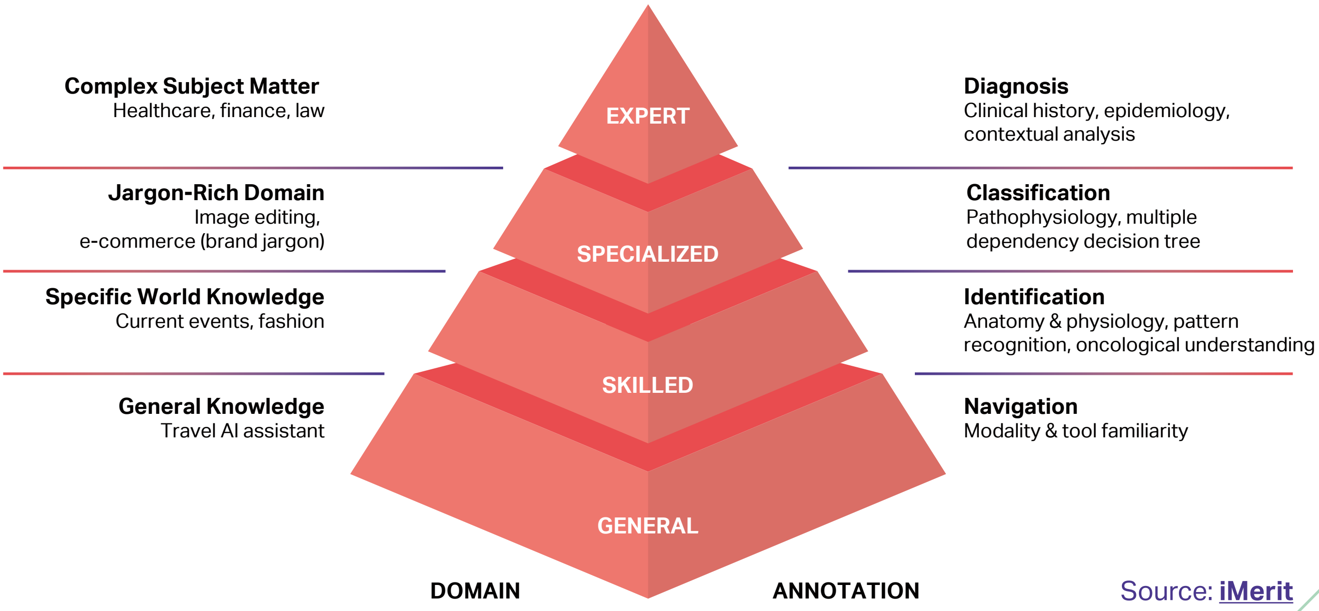
In addition to pattern recognition skills, prospective annotators must demonstrate general knowledge and curiosity about their environment. This is not to be confused with a high degree of schooling. There is a vast population that lacks educational resources, or has access only to inferior educational resources, but possesses all the natural intelligence necessary to become domain specialists with the proper narrow-but-deep training. Identifying and training these individuals is one of iMerit's unique value propositions. For our medical annotation specialists, meticulous and precise aptitude in pattern recognition is married to a traditional curriculum in anatomy, physiology, and disease processes.

CURRICULUM DESIGN

The curriculum begins with learning the toolset, which gets more advanced on a regular basis, from 2D images to multi-planar navigation in 3D imaging to 4D CINE studies. And while broad medical training is not required, the narrow-but-deep curriculum does include anatomy and physiology of the relevant area(s) for the use case under examination, and an understanding of the terminology as well. This elevates the annotation work from simple recall to deeper understanding creating a more robust skillset. For example, teams working on cardiology projects can understand anatomy and physiology well enough to commonly navigate congenital and pathologic aberrations. Because all training is hands-on, rather than based solely on textbooks and lecture halls, progress is typically quick, and getting quicker as pedagogical techniques are refined. This is a very agile microskilling model that allows people to learn and relearn new use cases, often within one-week cycles.

The modular curriculum is also technology-centric, and includes gamification, anywhere/anytime learning, and a high degree of personalization. For example, a custom-designed tool to assess the trainee's computer vision skills quickly identifies strengths and weaknesses and allows for additional focused training. A result of iMerit's social impact model is that motivation levels are very high and attrition levels are very low. Jobs at iMerit are life-changing so the investment in Learning & Development and the accumulation of knowledge and skill over time, is retained in the organization and customers love working with stable teams. Clients report that they like investing in training iMerit's teams because they see the fruit of their efforts over time.

Domain expertise meets annotation expertise



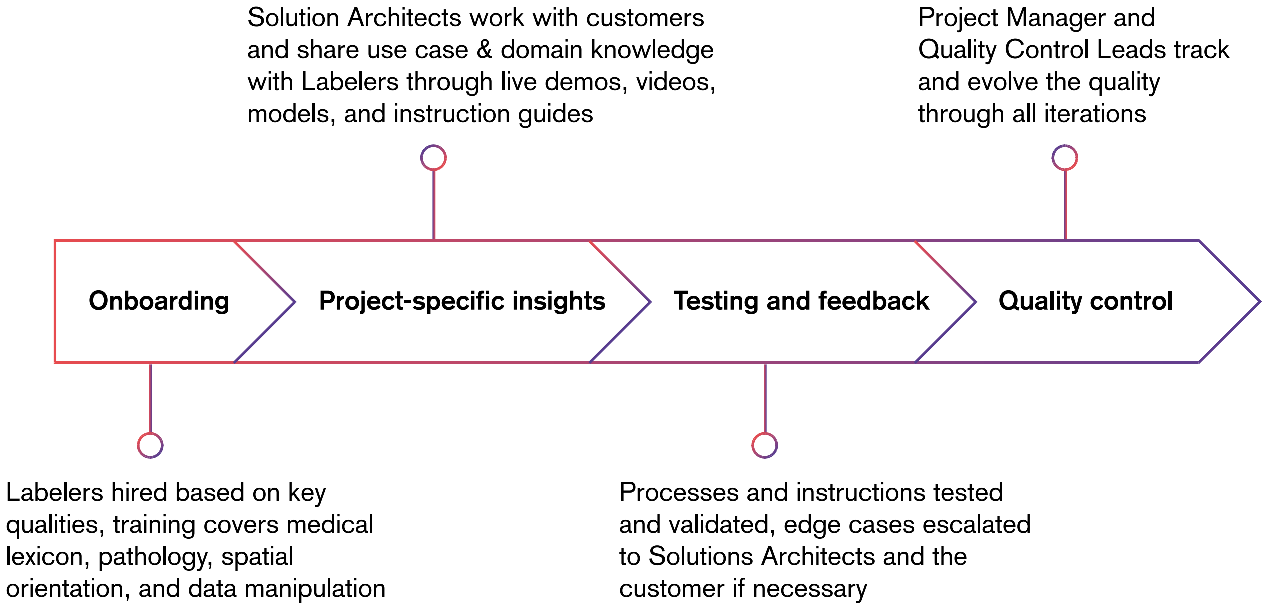
Source: iMerit

HUMAN FEEDBACK

With regard to machine learning it is a general rule of thumb that “more is more”: the more data that can be input into the algorithm, the more efficiently and accurately that algorithm will perform – often conferring a competitive advantage when an end product reaches market. However, even with massive amounts of data, those levels of accuracy will top out at around 60% to 65%. These levels are not acceptable, and serve to fuel the skepticism of medical AI that we hear in some quarters. To push those accuracy levels up to where they need to be, around 90%-95%, requires not more data but more human input. This is where human-in-the-loop feedback comes in. The data is re-examined by experts with more specialized training, such as radiology or histology technicians, nurses, or clinicians through iMerit’s expert network as needs are determined. By calibrating the exact balance of expertise and efficiency necessary for a project, we’re able to optimize the process for both cost and quality. When the data is then resubmitted to the algorithm, much higher levels of accuracy are achieved. It is important to note that iMerit skilled experts are available at all times for consultation by the non-expert specialist workforce. In fact, annotators are encouraged to seek guidance whenever they encounter an edge case.

It is evident, then, that the iterative processes behind machine learning, often thought of as machine-dependent alone, in reality demand human interactions which can materially affect a project’s timelines and costs in order to achieve the desired levels of accuracy.

How iMerit creates healthcare data labeling experts



Onboarding

Labelers hired based on key qualities, training covers medical lexicon, pathology, spatial orientation, and data manipulation

Project-specific insights

Solution Architects work with customers and share use case & domain knowledge with Labelers through live demos, videos, models, and instruction guides

Testing and feedback

Processes and instructions tested and validated, edge cases escalated to Solutions Architects and the customer if necessary

Quality control

Project Manager and Quality Control Leads track and evolve the quality through all iterations

For this reason human-in-the-loop actually begins at the earliest discussions of client engagement, and each project is customized through close consultation with the client. Before any teams can be deployed, both client and iMerit must be clear on the expertise requirements of the use case to assure the most cost-effective, efficient, and accurate outcome. As the old adage goes, there's nothing more expensive than doing it twice, both in terms of project budget and more importantly, time. Early investment in a consultative partnership, therefore, can often pay both immediate and long-term dividends.

SOCIAL SIGNIFICANCE

Throughout history and into our current era, remote villages have organized and prospered around developing marketable skill sets. Weaving, pottery, woodworking, metalworking, and beadwork are among the specialized crafts that have been unique to tribes and villages sometimes for centuries. This ancient practice has been repurposed by iMerit to bring value and prosperity to underserved communities through technology, realizing benefits both for the company, for its clients, and for the future of tech as a whole. Unlike the crowdsourced data annotators so ubiquitous in the AI realm, iMerit's annotators are full-time employees enjoying good wages and benefits, even during the training period, with many opportunities for advancement. In addition, it is a youthful workforce, one which was "born digital." The majority of these workers are women, who in general demonstrate a greater aptitude for annotative work. It is a model that iMerit has been able to port to new venues in several countries, including the United States.

THE BOTTOM LINE

iMerit is a for-profit business with a social impact mission, and its focus is always on its clients. The company's mission is to take AI from lab to production through the creation of a dedicated workforce. For medical AI the opportunities are just beginning. In addition to digital radiology, which has proven to be a clear starting point, iMerit is increasing its capacity to contribute human-in-the-loop capabilities for AI in:

- Digital Pathology
- Robotic Surgery
- Electronic Medical Records
- Telemedicine
- Pharmaceutical Research

As the field develops, iMerit will continue to pursue strategies assuring clients of:

- **Scalability:** Each engagement is the product of a detailed consultative process and custom designed for the client's specific use case. In the course of these consultations iMerit's medical experts design the curriculum for training the workforce. In addition, work volumes (i.e., the number of images to be annotated) can change dramatically from month to month. This means the workforce must be agile enough to adjust to changes in volume quickly and economically.

- **Accuracy:** The iMerit process has demonstrated proven results, achieving high accuracy without the necessity of re-annotating images several times, as is frequently the case when crowdsourcing annotators. The iMerit team includes skilled, medically trained supervisors to create and support training material, and review difficult edge cases as they occur.

- **Efficiency:** The preceding assure that the project is addressed in a way to achieve the greatest efficiency possible, at an economical cost. As AI becomes more ubiquitous for medical imaging applications, iMerit anticipates even greater efficiencies will be possible through the streamlining of iterative steps without sacrificing data quality.

- **Security:** iMerit annotators are full-time employees bound by non-disclosure agreements, with good salaries, benefits, and opportunities for advancement. Consequently, the company enjoys an employee attrition rate of less than four percent. For clients, high employee satisfaction translates into the knowledge that their proprietary data are secure.

- **Flexibility:** The annotated data provided by iMerit is tool- and platform-independent, so it can seamlessly and quickly be integrated into any ongoing project.

iMerit helps some of the world's largest companies deploy artificial intelligence accurately, securely, and with operational efficiency, offering high levels of data accuracy and lower effective costs to you for the data we process. We look forward to sitting down with you and designing a workflow plan that will move your project into actual production in a timely and cost-efficient manner.

Copyright © 2020 iMerit Inc. All rights reserved.

iMerit

THE CHALLENGE OF MED AI ANNOTATION

imerit.net