Google DeepMind's and IBM Watson's Advances into the Health Industry

Elizabeth Potapova

Department of Computer Science, Binghamton University

CS 301: Ethical, Social, and Global Issues in Computing

Dr. George Weinschenk

May 11th, 2022

Abstract

As artificial intelligence (AI) systems expand, they are applied to significant fields such as medicine. Two medical AI models are DeepMind's AI system (DMAIS) and Watson for Oncology (WFO), both of which analyze breast cancer. DMAIS impacts a wider population and is overall more adequate in comparison with WFO. DMAIS is a deep learning architecture made of a lesion, breast, and case model—the system detects breast cancer within a patient. WFO utilizes a question-answering system, where a question is posed in natural language and WFO uses a series of algorithms to determine the answer—the model recommends treatments for a patient already diagnosed with breast cancer. Both systems match the performance of professional radiologists. DeepMind recognizes the importance of reproducibility in research and publishes their work in detail. On the other hand, IBM has not published any specifics about WFO. Additionally, these AI systems need to be applicable to real-world scenarios, and DMAIS demonstrates that by being accurate when switching from one country to another. A well-crafted AI model must solve an urgent issue in the real world. Thus, DeepMind's AI system is more credible, applicable, and necessary compared to WFO. Computer science professionals should build AI systems that are needed, reasonable, and accurate. Professionals, such as doctors, should not rely solely on what an AI model may output, as such choices can drastically and negatively impact patients. Lastly, citizens should be aware that a combination of AI and professionals will soon be in charge of major aspects of society.

Google DeepMind’s and IBM Watson’s Advances into the Health Industry

Within our technical lives, people encounter artificial intelligence in various services: mobile phone assistants, auto-composition for emails, and social media feeds. The field of artificial intelligence (AI) has progressed in the past decade, with the increase in computational power allowing AI models to learn and process things within a reasonable time frame. Once AI systems were practical for major applications, two researcher groups focused on employing them in the health industry: DeepMind and IBM. Of many fields in medicine, DeepMind and IBM both decided to produce algorithms for breast cancer detection. DeepMind's AI system focuses on detecting breast cancer and can impact a wider population, compared to IBM’s Watson for Oncology that only recommends treatment to patients already diagnosed with cancer and thus has limited use. To understand the differences between the two algorithms, one must first explore the technical details of both Watson for Oncology and DeepMind’s AI system. Then, the accuracies of the two systems can be compared as they were both evaluated against a board of professionals. Finally, the development of these types of systems creates a social impact that requires analyzing: the credibility of the researchers designing AI systems, if such systems can be applied in the real world, and if these AI systems are necessary. Of the two research groups, IBM was the first to develop AI for medicine.

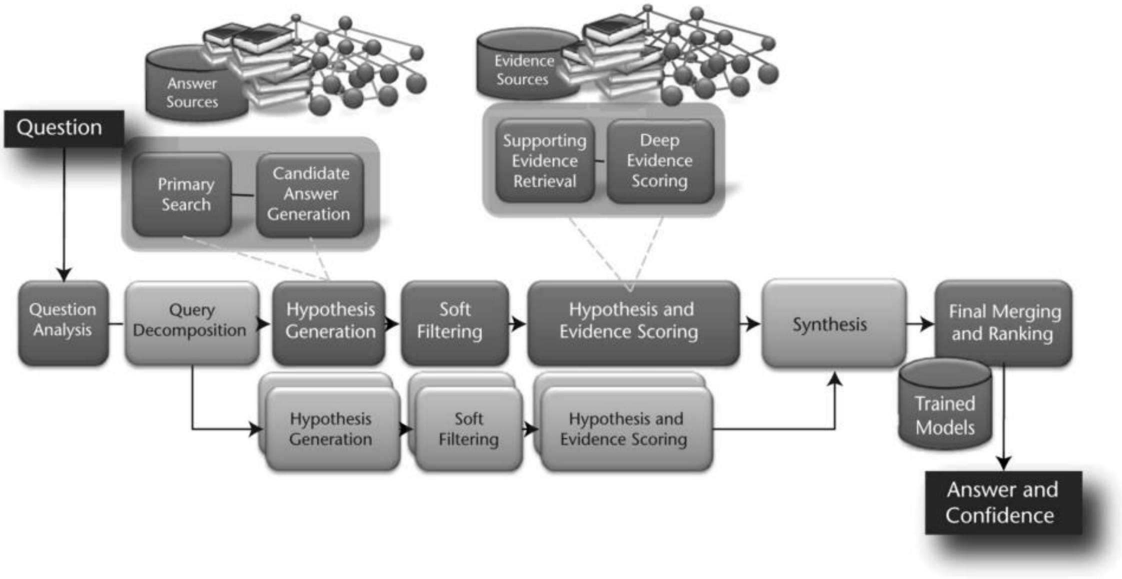
IBM’s Watson

IBM’s AI program, named Watson, did not start with a goal such as analyzing breast cancer: it was first developed as a tool to play *Jeopardy!*, the television quiz show (IBM, 2018, para. 1). After the algorithm’s success in *Jeopardy!*, IBM expanded the program into different fields, creating services like Watson Media, Watson Works, and Watson Health (Jennings, 2011, para. 8). One of the first explorations of Watson Health was Watson for Oncology (WFO), to help in the study of cancer (IBM India, 2016, para. 1).

WFO had the same underlying algorithm as did Watson for *Jeopardy!*: a question-answering architecture, where a question is proposed by a human in natural language and Watson provides an answer (IBM, 2014, para. 9). Due to this methodology, Watson is a general algorithm that depends on the question and what data is provided (Ferrucci et al., 2010). Under the hood, Watson combines question-answering techniques with deep neural networks into an algorithm dubbed DeepQA (IBM, 2014, para. 4). David Ferrucci (2010), lead of the Watson research team, describes the process of the algorithm: analyzing the question, hypothesizing the correct answer, scoring those hypotheses, and outputting one final answer based on which hypothesis has the highest score (see Figure 1) (p. 69).

Figure 1

General Process of IBM's DeepQA Algorithm



Ferrucci (2010) explains that prior to any questions or answers, the program must create a collection of sources that it will look parse throughout the entire process (p. 69). A series of documents are provided by the users of the algorithm, and DeepQA searches the web for new sources that are relevant to those documents, Ferrucci continues. DeepQA now has a large collection of sources to use during the hypothesis creation and scoring parts of the process.

With the collection of data happening prior, DeepQA is ready to process questions that are asked by the user. When DeepQA is presented with a question, the system uses “many independently developed answer-typing algorithms” to deduce what the question is asking (Ferrucci et al., 2010, p. 70). Ferrucci explains that the algorithms cover a variety of topics: one algorithm focuses on analyzing a question to see if it is asking for a location, and a separate algorithm checks for a person. The goal is to maintain each independent algorithm’s strengths while providing DeepQA with flexibility in the questions it can answer. These algorithms all work in tandem to process the question posed in natural language.

After determining what the question is asking, DeepQA generates possible answers that could answer the question. The model searches its collection of sources to find relevant documents, and then extracts concrete words or phrases from those documents that match the answer format (p. 71). Ferrucci affirms that this part of the process is the most crucial: if DeepQA does not produce the correct answer as one of the hundred candidate answers, then the question cannot be answered (p. 72). The snippets of words or phrases make up a collection of potential answers to the posed question.

With the set of possible answers, DeepQA scores each one based on evidence found in various sources. DeepQA consists of multiple smaller scoring algorithms that each calculate a separate score for how well a piece of evidence matches the candidate answer (p. 72). Then, a culmination of those scores produces one metric for how accurately a candidate answer matches the question. Thus, the set of candidate answers now includes one value that measures how well each piece of evidence of an answer matches the question.

Finally, DeepQA uses various trained neural network models to rank each candidate answer (p. 74). As Ferrucci describes it, “certain scores that may be crucial to identifying the correct answer for a factoid question may not be as useful on puzzle questions.” The hypothesis with the highest rank is outputted as the answer to the question. Through this process is DeepQA able to provide an answer to a question posed in natural language.

IBM named its DeepQA algorithm as Watson, in honor of IBM’s first CEO Thomas J. Watson (Hale, 2013, para. 2). Instead of being provided with trivia resources, WFO had access to past breast cancer cases and patient information, such as doctor’s notes, symptoms, and tumor screenings (Somashekhar et al., 2018, p. 419). WFO is an algorithm built on robust question-answering architecture that can recommend treatment options to doctors. On the other hand, DeepMind’s AI system can determine if cancer is present in a patient or not.

DeepMind’s Artificial Intelligence System

DeepMind’s artificial intelligence system (hereafter, DMAIS) is a deep learning model. The architecture consists of three individual models whose outputs, each a confidence score from 0 to 1, are combined to give one final evaluation (McKinney et al., 2020b, p. 96). A confidence score is also referred to as a prediction score or a cancer risk score; the generated 0 to 1 can be seen as a 0% to 100% chance of the case containing cancer. Each model receives four mammography images (referred to as the case): one top-down view and one profile view of the breast for each breast (p. 95). These models analyze the images differently: the lesion model examines individual tumors and damaged tissue, the breast model examines one individual breast at a time, and the case model examines both breasts simultaneously (p. 96). The paper specifies that the models analyze all four provided mammograms, but they provide confidence scores for each area as specified before aggregating them into one confidence score for that model. Each model utilizes image manipulation techniques that are applied to the photographs; these methods increase accuracy, an invaluable goal of AI in medicine.

Image Manipulation

Deep learning requires for a lot of data in order to be accurate, but large quantities of data are hard to come by, especially in the medical field (Shorten & Khoshgoftaar, 2019, p. 4). In order to bypass this problem, deep learning models often use data augmentation, specifically in regard to images, to generate a larger dataset for the model to learn from (p. 3). Instead of creating new images, such models modify existing ones using techniques like stretching or flipping. To a neural network, the resulting image appears to be an entirely new one.

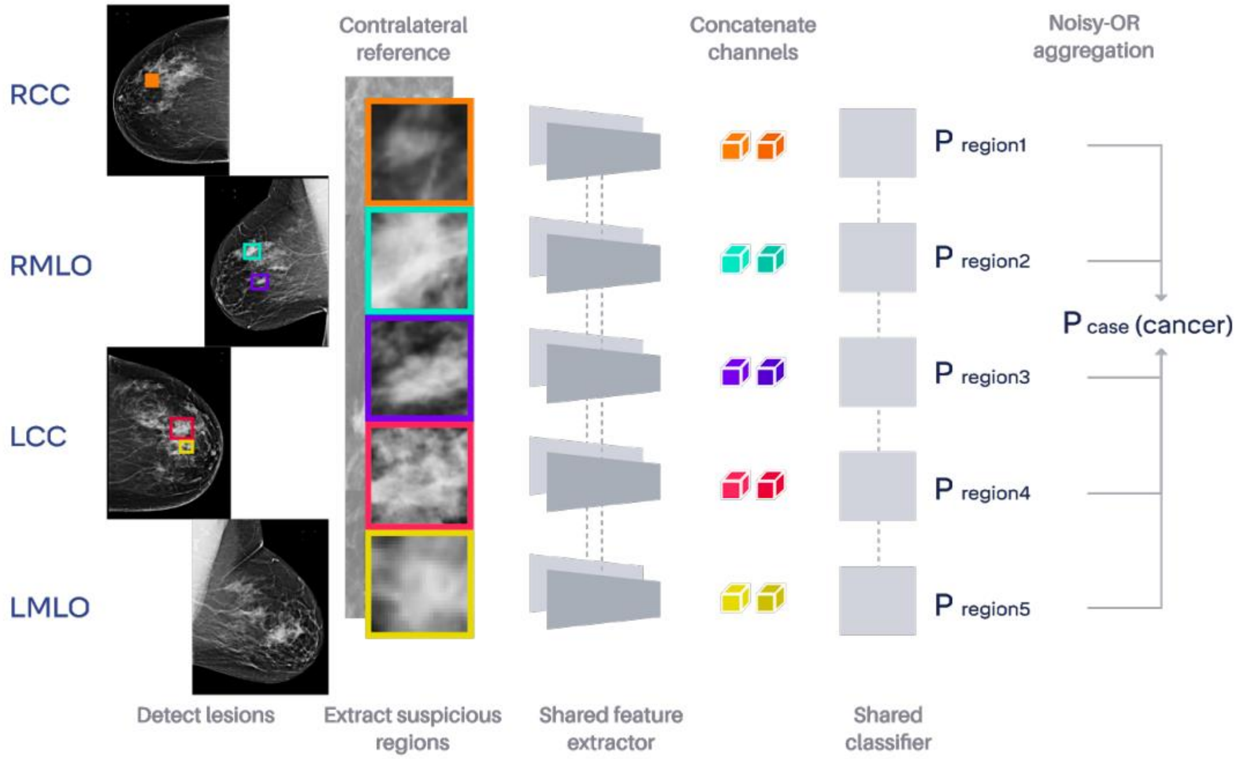
DMAIS applies data augmentation to all testing images within its three separate models, including but not limited to “elastic deformation, shearing, rescaling, translation, and flipping” (McKinney et al., 2020a, Suppl. p. 1). Although there are some slight variations in layers and parameters, all three models had the same methods of manipulating input images for training. Applied in order, all models typically had elastic deformation by a random magnitude, rotation by a random angle, horizontal and vertical scaling up to ±10%, 50% chance of horizontal and vertical flipping, horizontal and vertical shearing by a magnitude within ±10%, horizontal and vertical translation by a random pixel value, and cropping (Suppl. pp. 2-4). By applying data augmentation techniques to the mammography images, each separate model has a larger dataset to train on. This results in a higher accuracy as the models learn to identify lesions in a variety of differing positions throughout the breast.

Lesion Model

One of the models of DMAIS consists of three steps: identifying all possible lesions, producing a confidence score for each lesion, and then combining all individual lesion scores to output final confidence score for the case (see Figure 2) (McKinney et al., 2020a, Suppl. p. 7).

Figure 2

Architecture of DMAIS' Lesion Model



Note. Abbreviations on left refer to mammography image angle: CC refers to craniocaudal view (top-down angle), MLO refers to mediolateral oblique view (side angle, as seen from center of chest outwards), and R and L specify right and left breast, respectively. P refers to the prediction score (how likely cancer is present in the region or case).

Given one image from the case, the model uses RetinaNet to identify areas that could be tumors or other cancerous tissue (Suppl. p. 1). RetinaNet is a classifier, which use convolutional neural networks to extract features from images; it is an algorithm designed by Lin et al. (2020) that DMAIS then utilizes. After analyzing all four images in the case, RetinaNet produces multiple small images that are cropped to the cancerous lesions (McKinney et al., 2020a, Suppl. p. 1). Those images are augmented based on the parameters stated earlier (Suppl. p. 2). The paper continues, stating that those regions then have to be located in the opposite view of the breast: for example, if a possible lesion is located in the profile view of the breast, the model has to pinpoint the same lesion in the top-down view. As per McKinney et al., the cropped images and broad regions of the opposite view of the breast are both passed to MobileNetV2. This algorithm, created by Sandler et al. (2018), is a convolutional neural network that extracts shared features between two images. The combination of the two different networks produces one lesion at a specific location.

After individual regions are identified, the lesion model then scores each one based on how likely they are to be cancerous tissue. McKinney et al. (2020a) create and train a model to produce such scores; to create an accurate algorithm, it is trained with "positive and negative examples [of lesions] in equal proportion," and then tested on images that had image manipulation applied identical to the training cases (Suppl. p. 2). With a compete scoring model, all individual regions are passed through the model to produce a collection of region cancer risk scores.

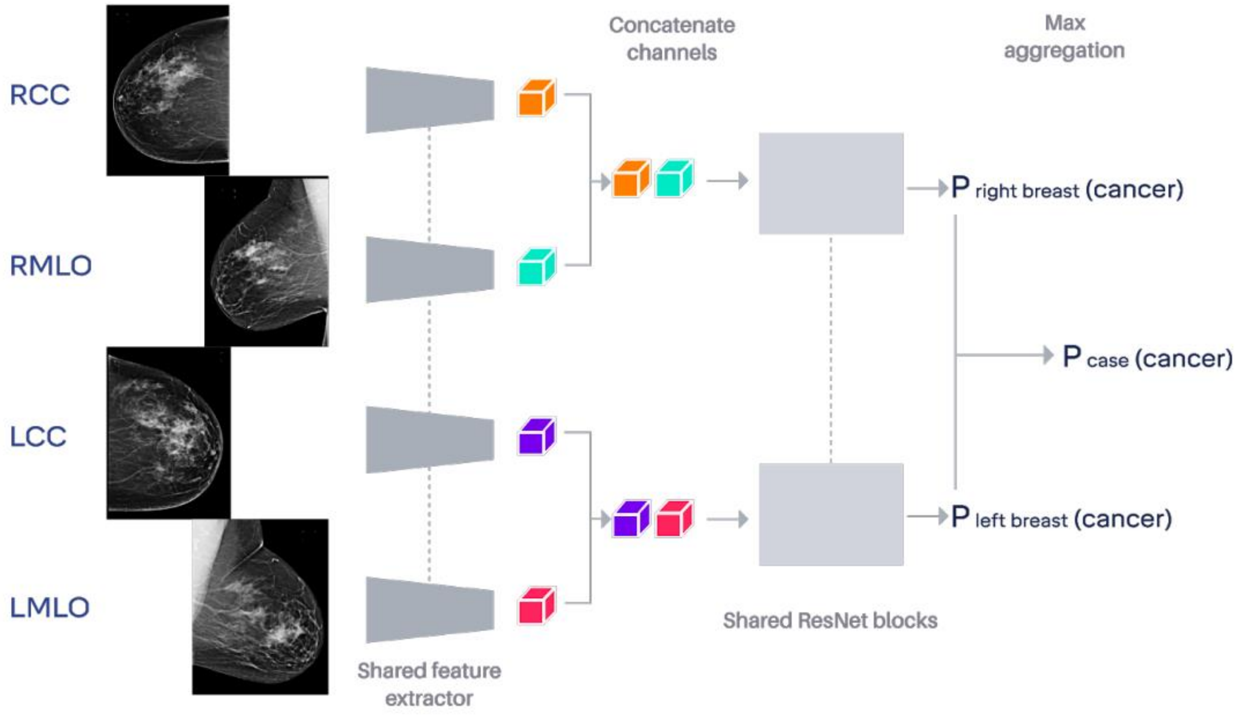
Finally, the individual scores are combined into one confidence score representing the entire case. They are merged "using the noisy-OR operation" (McKinney et al., 2020a, Suppl. p. 2), which can be described as a "generalization of the logical OR" (Blutner, 2012, p. 2). If at least one region produces a score stating that it contains a hazardous lesion, then the cancer risk score of the entire case would reflect that individual score.

Breast Model

The other two models, the breast and case model, run simultaneously and separately from the lesion model. The breast model of DMAIS examines each individual breast for cancerous features before outputting a confidence score for the mammography case (see Figure 3) (McKinney et al., 2020a, Suppl. p. 7).

Figure 3

Architecture of DMAIS' Breast Model



Note. Abbreviations on left refer to mammography image angle: CC refers to craniocaudal view (top-down angle), MLO refers to mediolateral oblique view (side angle, as seen from center of chest outwards), and R and L specify right and left breast, respectively. P refers to the prediction score (how likely cancer is present in either breast or the case).

Each image is modified with the image manipulation parameters stated above (McKinney et al., 2020a, Suppl. p. 3). Then, features such as potential lesions are extracted from the image using

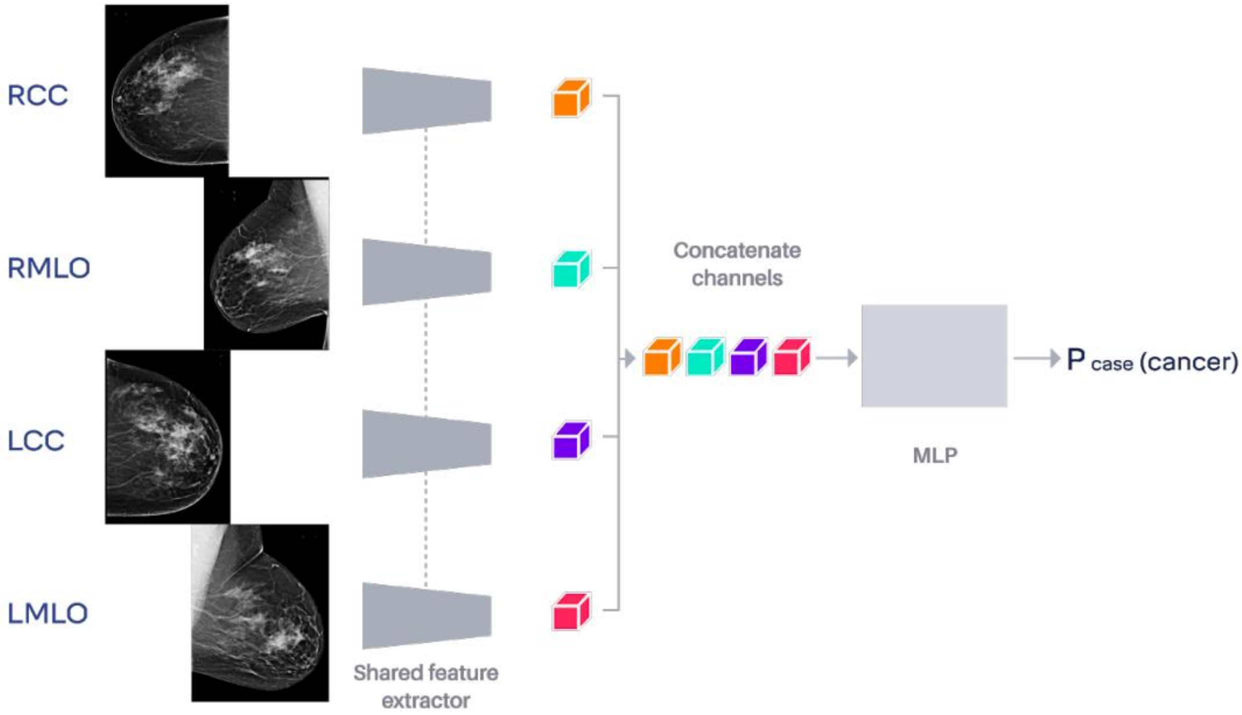
ResNet-v2-50, an algorithm explored by He et al. (2016b). Extracted regions are linked together and passed through an additional neural network that deduces a breast-level confidence score (McKinney et al., 2020a, Suppl. p. 3). Finally, McKinney et al. state that the maximum score of the two breasts is outputted by the model as the cancer risk score for the case.

Case Model

Alongside the lesion and breast model, the case model analyzes all four mammography images together (see Figure 4) (McKinney et al., 2020a, Suppl. p. 7).

Figure 4

Architecture of DMAIS' Case Model



Note. Abbreviations on left refer to mammography image angle: CC refers to craniocaudal view (top-down angle), MLO refers to mediolateral oblique view (side angle, as seen from center of chest outwards), and R and L specify right and left breast, respectively. P refers to the prediction score (how likely cancer is present in the case).

First, the images are augmented with modifications stated previously, with an additional layer having a 50% chance of swapping the images of the left and right breasts (McKinney et al., 2020a, Suppl. pp. 3-4). Features are extracted from each image equally using the ResNet-v1-50 algorithm (Suppl. p. 4), created by He et al. (2016a). Linked together, the regions are passed through a hidden layer of 512 nodes to evaluate the confidence score for the entire case (McKinney et al., 2020a, Suppl. p. 4). This value is then what the case model outputs to DMAIS.

Combined Architecture

Given a breast cancer screening case, DMAIS passes the images to the lesion, breast, and case models which evaluate them separately and in no particular order. The final score for the entire case is the mean of the three individual model outputs, with all scores ranging from 0 to 1 to signify the likelihood of cancer (McKinney et al., 2020b, p. 96).

Comparing Results of WFO and DMAIS

After each separate model is designed and trained, they must be tested for accuracy before being applicable in hospitals. Both WFO and DMAIS are evaluated by comparing their outputs with those of a board of radiologists (McKinney et al., 2020b, p. 91; Somashekhar et al., 2018, p. 419). Identical cases are passed to both the AI system and the board, and the system is considered accurate if its output matches the evaluations of the experts (McKinney et al., 2020b, p. 91; Somashekhar et al., 2018, p. 419).

The evaluation for WFO had a board of 15 members for 638 cancer cases from India (Somashekhar et al., 2018, p. 419). Both the board and WFO had access to the patient’s clinical trials and reactions to past treatments (p. 419). Treatment recommendations provided by WFO matched with those of the tumor board an overall 93% of the time, with slight variation based on the stage of cancer (p. 420).

DeepMind's system had 25,856 cases from the UK and 3,097 cases from the US for a total of 28,953 cases (McKinney et al., 2020b, p. 90). After the model was trained, it was evaluated retrospectively by examining if the system's output matched the historical reviews of the radiologists provided in the dataset (p. 90). For both the UK and US datasets, McKinney et al. states that DMAIS matched the performance of the specialists with a 5% margin, and in some cases outperformed them.

WFO recommends treatments for patients diagnosed with breast cancer, whereas DMAIS identifies such cancer within women. Although the difference in datasets seems drastic and incomparable, the number of mammography cases resulting in a positive diagnosis was 1,100 of the total 28,953 cases (McKinney et al., 2020b, p. 90). When put into perspective, WFO's dataset of 638 cases is reasonable to compared to that of DMAIS's. In terms of sheer data, DMAIS would provide more accurate results, as it was trained on more cases from multiple countries compared to WFO. Although both WFO and DMAIS produce results on par with professionals, the difference in the systems comes down to how they impact society. The companies behind the AI models have diverging approaches on conducting and sharing research.

Social Impact

Credibility

Reproducibility in research leads to credibility. A paper's conclusions would not be considered trustworthy if a separate research group is unable to reach identical results. Additionally, reproducibility allows for a researcher to build upon another author's work: new discoveries are built upon the results of old ones. This idea of reproducing research is a parallel to open-source software. Brad Griffith (2016), an entrepreneur and engineer, describes open-source as "understanding what you're really good at and … putting your best work out there for people to see and criticize and build upon" (5:01). Recognizing weaknesses is applicable to both research and software development; combining strengths in a collaborative effort yields more accurate and quality results.

DeepMind recognizes the importance of reproducibility. After publishing their paper on DMAIS, Haibe-Kains et al. (2020) criticized DeepMind for the lack of transparency in their methods, specifically the missing parameters needed to recreate the deep learning models used in DMAIS (p. E14). In response, McKinney et al. (2020a) updated their descriptions of the models and specified how other networks (such as ResNet and MobileNet) were implemented. The quick response and open-mindedness exhibit DeepMind's values in allowing other researchers to reproduce and build upon their work.

In contrast, IBM maintains a tight grasp on the inner workings of their AI models. To note, the previous analysis of DeepQA, the underlying architecture of WFO, was based upon papers published a decade ago on Watson for *Jeopardy!*. Although WFO works on a similar structure, IBM has not published peer-reviewed research articles describing exactly how the algorithm works. When viewing WFO alongside IBM's other goals for Watson, such as Watson Advertising and Watson Media, it is clear that WFO is one of many in a line of products. DeepMind is developing AI models to expand the field of medical AI and allow for other researchers to build upon their work; IBM is expanding Watson to create a new product that makes a profit.

Applicability

DMAIS and WFO are not the first AI models to be implemented in the medical field. In an article commenting on DeepMind's work, Etta Pisano (2020), a radiologist and researcher, mentions computer-aided detection (CAD) systems that were developed to assist in medical imaging at hospitals (pp. 35-36). She stated that CAD “showed great promise in experimental testing but fell short in real-world settings” as a warning for AI models that exhibit amazing results in controlled settings (p. 35). Pisano mentions possible causes for the failure of CAD: inability to compare screenings with past ones of the same patient, misuse by radiologists, and the mental approach of specialists using CAD (p. 36). Due to limited processing power, Pisano explains that CAD would “mark regions that were not changing over time and that could be easily dismissed by expert readers.” Additionally, she mentions that if an area was marked by CAD but the radiologist saw no sign of cancer, the specialist would mark the system for error instead of investigating the region. Lehman et al. (2015) examined the use of CAD throughout 66 hospitals and found that CAD did not improve radiologist accuracy and decreased an expert’s sensitivity to cancer present in a screening (p. 1828). Although relatively small, these small issues with CAD would lead to an increase in false negatives, a danger to patients. Perhaps the radiologists’ mentality toward CAD affected its real-world performance: if the expert distrusts the system or does not understand how it reached its conclusion, CAD’s outputs would be ignored, leading to missed cases of cancer. On the other hand, reliance on such a system would lead to the dulling of the expert’s skills. CAD systems are a prime example of what new AI models should be aware of: explaining output to experts, being flexible with its use, and establishing trust.

The failure of CAD provides insight into what is needed for a realistic AI system. Although DMAIS is far from real-world application, it is in a better position than WFO. DeepMind’s architecture has proven to be highly accurate when switching from one country to another (McKinney et al., 2020b, p. 91). Within its separate models, individual regions are marked and paired with the same lesions in different perspectives. Expanding the techniques already in place in order to show radiologists DMAIS' process would not be an extremely difficult task. When talking about explainable AI, Storey et al. (2022) mention that understandable reasoning for an AI's decision is crucial "to ensure decision making is justified, fair, and ethical, and to treat the 'right to explanation' as a basic human right" (p. 27). The stated "right to explanation" could aid healthcare professionals as well as patients who may not immediately trust an AI system with their diagnosis. Additionally, DeepMind's values on reproducibility allows for other researchers to contribute in small but specific areas that they are experts in. With DMAIS having a strong core AI system, explaining how the system reached its conclusions and establishing trust within the radiologist community would be tedious but not impossible.

Necessity

When discussing the applications of medical AI, is it important to recognize where such models are necessary. If the production of a system costs millions of dollars and provides no significant benefit, then the development of such a system is not needed. Although WFO is an impressive feat for question-answering algorithms, it is designed solely to provide possible treatments for patients already diagnosed with breast cancer. On the other hand, DMAIS's detection of cancer leads to said diagnoses and treatments. When evaluating which model would be more valuable in the medical field, one must consider current problems in the world.

There exists a disproportionate ratio of health care providers to patients. Simply put, there are not enough specialists for the number of patients, especially in fields where regular screenings are required, like breast cancer (Peng, 2020, 3:38). Lily Peng (2020), a doctor and current project manager, describes her work at deploying an AI system at Google Health. The project focuses on eye disease caused by diabetes, and how AI can process screenings to help catch early detection of such diseases in India (4:35). Peng mentions how there are roughly 15,000 eye specialists for India’s 62 million diabetic patients (3:38). Although not every patient develops complications, yearly screenings are recommended in order to prevent blindness (3:14). Thus, the number of screenings that the small amount of doctors have to sift through is insurmountable. An AI system, such as Peng’s team is developing, can process yearly screenings so that eye doctors can focus on providing better care to patients already diagnosed with complications. Peng’s system

The question comes to whether DMAIS or WFO are needed in modern medicine. DMAIS tackles the same issue of processing regular breast cancer screenings so that doctors can provide their patients with the best healthcare. On the other hand, WFO can only be applied to patients who have already been diagnosed. Modern problems lie not in being unable to decide on proper treatments but recognizing when a patient has complications and getting them to that appointment. Thus, DMAIS would provide more utility to the medical field in comparison to WFO.

Summary

DeepMind’s system is not only more useful than WFO, but more adequate in a variety of factors. DMAIS is made of deep learning architectures and WFO utilizes a question-answering system, but both systems perform on par with medical professionals. DeepMind recognizes the importance of reproducibility in research and publishes their work in detail. On the other hand, IBM has not published anything about WFO specifically, with the only information about its inner workings being released over a decade ago. Additionally, these AI systems need to be applicable to real-world scenarios, and DMAIS demonstrates that by being accurate when switching from one country to another. Lastly, a well-crafted AI model may not be necessary in the real world—one must recognize which system provides an immediate effect. DMAIS detects if a patient has cancer or not, whereas WFO simply provides options of treatment to patients already diagnosed. Thus, DeepMind's AI system realistically aids patients and impacts a broad population in comparison to WFO.

DeepMind's model symbolizes the growth of practical AI and how AI systems can have a major impact around the world. Computer science professionals, especially those in research, should determine if the systems they develop are necessary and relevant to practical situations. For AI systems, accuracy in the output and testing on realistic data are a must. Finally, explanations of the model's output are obligatory for a system used in vital fields such as medicine. With artificial intelligence systems surrounding everyday life, people rely on them being used reasonably and accurately. Professionals should not rely solely on an AI model, and citizens should not be subject to poor outcomes due to such choices. As advancements in computers continue, people need to be aware of the permanent joint effort of artificial intelligence and professionals within vital parts of society.

References

Blutner, R. (2012). *Noisy OR* [Slides]. Blutner.De. http://www.blutner.de/Intension/

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010, July 28). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*(3). https://doi.org/10.1609/aimag.v31i3.2303

Griffith, B. [TEDx Talks]. (2016, May 18). *How open-source software can change our lives | Brad Griffith | TEDxNewAlbany* [Video]. YouTube. https://www.youtube.com/watch?v=hFRS46PsDU0

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Shraddha, T., Kusko, R., Sansone, S. A., Tong, W., Wolfinger, R. D., Mason, C. E., Jones, W., Dopazo, J., Furlanello, C., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., . . . Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16. https://doi.org/10.1038/s41586-020-2766-y

Hale, M. (2013, February 5). *"Nova" looks at I.B.M. computer that plays "Jeopardy!"* The New York Times. https://www.nytimes.com/2011/02/09/arts/television/09nova.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. *arXiv*. https://doi.org/10.48550/arXiv.1603.05027

IBM. (2014). *The DeepQA research team: DeepQA*. IBM Research. https://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2159

IBM. (2018). *A computer called Watson*. https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/

IBM India. (2016, July 29). *Manipal hospitals announces national launch of IBM Watson for Oncology*. IBM. https://www03.ibm.com/press/in/en/pressrelease/50290.wss

Jennings, K. (2011, February 17). *My puny human brain*. Slate Magazine. https://slate.com/culture/2011/02/watson-jeopardy-computer-ken-jennings-describes-what-it-s-like-to-play-against-a-machine.html

Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, *175*(11), 1828. https://doi.org/10.1001/jamainternmed.2015.5231

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. https://doi.org/10.1109/tpami.2018.2858826

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020a). Addendum: International evaluation of an AI system for breast cancer screening. *Nature*, *586*(7829), E19. https://doi.org/10.1038/s41586-020-2679-9

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020b). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Peng, L. [TEDx Talks]. (2020, June 25). *Democratizing healthcare with AI | Lily Peng | TEDxGateway* [Video]. YouTube. https://www.youtube.com/watch?v=MNp26DgKxOA

Pisano, E. D. (2020). AI shows promise for breast cancer screening. *Nature*, *577*(7788), 35–36. https://doi.org/10.1038/d41586-019-03822-8

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00474

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Somashekhar, S., Sepúlveda, M. J., Puglielli, S., Norden, A., Shortliffe, E., Rohit Kumar, C., Rauthan, A., Arun Kumar, N., Patil, P., Rhee, K., & Ramya, Y. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, *29*(2), 418–423. https://doi.org/10.1093/annonc/mdx781

Storey, V. C., Lukyanenko, R., Maass, W., & Parsons, J. (2022). Explainable AI. *Communications of the ACM*, *65*(4), 27–29. https://doi.org/10.1145/3490699