Machine Learning Performance Metrics and Diagnostic Context in Radiology

Henrik Strøm
Digital
Copenhagen School of Design and Technology
Copenhagen, Denmark
hstr@kea.dk

Steven Albury
Digital
Copenhagen School of Design and Technology
Copenhagen, Denmark
stal@kea.dk

Lene Tolstrup Sørensen
Center for Communication, Media, and Information Technologies
Aalborg University
Copenhagen, Denmark
ls@cmi.aau.dk

***Abstract*—In this pilot study data gathered from interviewing specialists in radiology is combined with an assessment of the way machine learning metrics are used in studies of radiological work. It argues that situated context of use should be an important contributor to the design of machine learning applications in radiology. The article shows how radiologists see their professional practice as utilizing a wider range of expert knowledge than many existing studies on machine learning in radiology allow for. The article describes a case study drawn from radiology practice in a major Danish hospital and discusses a widely cited study on machine learning in radiological work. The study connects current understandings of appropriate metrics used by machine learning researchers with professional radiologists' understanding of their diagnostic work. This comparison helps identify gaps in understanding between these two communities and suggests how they might be addressed.**

Keywords—machine learning, metrics, diagnostic context, radiology

I. Introduction

Machine learning has become an important field for the development of new types of computing systems and extending the types of activity that can be augmented by software. Google Deepmind's AlphaGo, IBM's Watson and advances in the development of autonomous vehicles [1] has raised the awareness of machine learning in public discourse. Recently, machine learning has raised headlines in healthcare. For example, Google has developed a machine learning algorithm to identify cancerous tumors on mammograms,[2] while at Stanford University, an algorithm is helping to identify skin cancer.[3]

Machine learning can be understood as an approach to computer science that allows computer programs to use large volumes of data and computational techniques to make decisions without human intervention and to 'learn' by adapting over time. Essentially, this means that machine learning algorithms become better with the quantity of data they can get for improvements. In this context 'most effective' means the simplest computational model that fits the success criteria for the problem being studied. The area of healthcare is known for large quantities of data and many manual processes which is one of the reasons why machine learning has been identified as a potentially beneficial technology for the healthcare sector.[3] For example, radiology of the thorax has been identified as one of the areas within healthcare where machine learning can be useful.[4] In Rajpurkar et al. [4] researchers applied machine learning to a number of x-rays and show how this can identify and classify pathologies. However, the paper also raises questions about work processes and the data needed for such work.

This paper focuses on analyzing the possibilities for applying machine learning in work processes in radiology by exploring the fit between common metrics and how well they meet the needs of practitioners in a real-world radiology setting.

The paper is built on a combination of desk study of widely cited literature and a case study where an interview with radiologists from Rigshospitalet, Denmark, was carried out to discuss work processes, and possibilities for using machine learning as part of the diagnostic practices of radiologists who specialize in the thorax (this encompasses the chest, lungs, etc.)

The remainder of this paper is organized as follows: we describe the methods used in section II. We then introduce the topic of machine learning in healthcare in section III. In section IV we introduce a case study we carried out in collaboration with Rigshospitalet in Copenhagen, Denmark. An analysis of our case study and relevant, related literature is presented in section V. Finally, section VI concludes this paper.

II. Methods

The study is a pilot study and raises questions that need to be followed up with further investigation. A larger dataset and a longer period of examination will allow for improved assessment of both qualitative factors and the impact of new technological developments on performance metrics. The study therefore follows a pragmatic, mixed methods approach where interview and document analysis are augmented by exploration of the appropriateness of widely cited approaches to performance measurement on a real-world environment.[4] This builds on other work in health informatics research and raises questions around study design and the importance of

978-1-7281-0444-7/18/$31.00 ©2018 European Union

innovation and workplace studies in healthcare.[5][6] This paper assumes that technological innovation intends to reshape practice to take advantage of the benefit claimed for the new technology. By combining an examination of the appropriateness of the measures being used in machine learning based innovation in radiology, with an exploration of the sociotechnical environment in the specific context of use (a large hospital) we argue that a more complete picture of the success of an innovation can be obtained. This is an approach that Greenhalgh and Swinglehurst [7] argue for in their discussion of how an increased use of ethnographically grounded research can help raise awareness of the potential problems of purely experimental, quantitative study designs. They argue that many such designs are technocentric, and make a lot of positivistic assumptions about the environment in which the innovation will be used. This focus on the technical elements of the research is something that this study found might lead to inappropriate measurements being applied, which conflict with the needs of real-world practitioners. In such situations, practitioners can find the technology to be a constraining factor rather than an enabling one.

In order to explore the problem space, a limited range of methods was adopted due to the constraints of time and resources in a pilot study. The intention was to trial the effectiveness of qualitative approaches, drawn from existing qualitative frameworks, including grounded theory and activity theory to provide a naturalistic perspective. By combining this with a critique of the quantitative metrics in an existing technical study of machine learning the project examined if there was evidence to suggest the context of use should be brought into the innovation process earlier than is commonly the case in health technology projects. This human-sensitive approach to the requirements for such systems has been found to be useful in large-scale sociotechnical systems requirements engineering, e.g. air traffic control systems.[8]

In order to gather the qualitative data, an interview was held in the office of two senior medical practitioners at Rigshospitalet. Both practitioners are highly experienced radiologists, and both were regular users of existing digital support systems used in radiology. In addition to the interview observation notes of the environment and movement of other practitioners in and around the workspace was taken. As discussed in section III an initial analysis by each member of the research team gave rise to two significant themes in the interview: the diagnostic context in which the diagnosis is made and the metrics used in measuring performance of the diagnostic process.

This classification process was the result of an initial sifting and then a closer analysis both individually by the research team and in joint sessions. Because of the relatively small amount of data gathered in a one hour interview, it was not felt necessary to use digital qualitative analysis tools at this stage. The analysis of quantitative metrics was carried out using a comparative approach to test the appropriateness of the classification and scoring measures used on radiology images with the needs of the practitioners. This is because for radiology work issues around individual patient history play a central role in decision making and it is clear from the paper selected for comparison that this is not fully taken account of in the design of machine learning-based image analysis.

In the following analysis the themes outlined above are discussed to develop the argument made in the conclusion that further study is needed, and that attention should be paid to the lived experience of radiologists and other practitioners when developing machine learning based tools.

III. Machine Learning in Healthcare

Recent advances in machine learning now enable us to solve problems that would otherwise be practically impossible to solve, by running machine learning models on very large datasets.[9] Using multi-layer neural networks with non-polynomial activation functions, it is possible to approximate any function, e.g., for classification and regression, given enough data.[10] These neural networks are able to learn from data that are labeled, e.g., classified as examples of presence or absence of the attribute used for classification, in a process called supervised learning. During this process, the neural network will slowly adjust its approximation to fit with the labeled data.

Researchers have produced impressive results,[11][12] and seemingly even outperforming specialist human performance in some fields, e.g., pneumonia detection in x-ray imagery.[4] However, such research focuses on a very narrow task, in this case, analysis of out-of-context x-ray imagery, that might not be directly applicable in a clinical setting.

As already mentioned, Google has developed machine learning algorithms which can identify metastases in breast cancer. In Strumpe et al. [2] it is described how the machine learning algorithms have been developed to detect metastases from breast cancer. The identification of metastases is essential for the patient and directly associated with the treatment the patient receives. The algorithm has been tested for robustness of data material by being applied to different datasets for improvement and accuracy.

At Stanford University, machine learning has been used for developing an application for detecting skin cancer. The work is described in Esteva et al. [13] where it is shown that their algorithms can detect both lethal as well as benign cancer from smartphone images. The application is expected to be further developed and by 2021 handle 6.3 billion smartphone subscriptions and in that way create a foundation for low-cost diagnoses.

Rajpurkar et al. constructed a 121-layer convolutional neural network and trained it with more than 100,000 samples of chest x-rays in a supervised learning setting, labeled by 14 pathologies for classification.[4] In a comparative test with experienced experts in radiology, the network surpassed the radiologists in diagnostic performance by the F1 metric.[4] The radiologists were not given access to patient records or any other information about the patients.[4] Andrew Ng,[14] a contributing author to the study, hinted that this might indicate that machine learning technology will soon be able to replace expert radiologists.

Due to digitalization in the healthcare sector, vast amounts of structured and unstructured data are now available in patient records and test results, making the introduction of machine learning in the healthcare sector "*inevitable*".[15] However, analyzing unstructured data such as prose notes in a patient record poses a major challenge, but technologies such as natural language processing can facilitate clinical decision support,[16] report classification,[17] and information extraction.[17]

Significant adoption of machine learning technology in clinical healthcare is yet to take place, as pathology classification from radiology imagery is still at an experimental stage.[4][11][12] Pons et al. make similar observations regarding the adoption of natural language processing for clinical decision support, report classification, and information extraction.[17]

The examples mentioned above illustrate the usefulness of machine learning in healthcare. However, the use of machine learning algorithms within healthcare can raise ethical concerns about which data material to use, false positives and data which are not used.[3] There is a need for researching in the area of healthcare to identify the possibilities as well as understanding the challenges and limitations of this approach.

IV. Case Study

Our case study was conducted as one observation session in parallel with a semi-structured interview with Chief Physician Martin Lundsgaard Hansen, MD, and Chief Physician Thomas Axelsen, MD, at the Department of Radiology, Rigshospitalet in Denmark, on September 13, 2018. The interviewees were both radiologists processing x-rays from patients with diseases of the thorax. Additionally, the radiologists provide input to the diagnosis of patients based on the x-rays and additional patient data. An audio recording of the interview was made. During this section, we will refer to this recording using timestamps. After the interview, the recording was partly transcribed, and single quotes and themes for the interview were identified. The interview was conducted in the interviewees' native language: Danish. In the following, quotes from the interview have been translated to English for purposes of transparency. The audio recording may be obtained from the authors of this paper by relevant third parties for research purposes.

The interviewees were introduced to the works of Osborne et al. [18] and Rajpurkar et al.,[4] along with Andrew Ng's tweet of November 16, 2017,[14] the latter as an illustration of how prominent machine learning researchers view the potential impact of machine learning on the field of radiology.

The interviewees described their work procedures and practices, and the opportunities and challenges they saw from introducing machine learning technologies into the field of radiology, and we asked questions for clarification. Because the machine learning technology we discussed is yet to be adopted by the field of radiology into clinical settings the interviewees (like other radiologists) have no experience in applying such technology in a clinical setting. Therefore, the interview took the study by Rajpurkar et al. [4] as a starting point for the conversation.

The remainder of this section will introduce our most significant findings from this interview. In addition to this, ethnographic field notes of the movement of other staff and the organization of work were taken to enrich the interview data.

There are two main deliveries from the radiology department, as the interviewees summarised:

- Raw radiology imagery that is delivered directly to third parties without prior analysis at the Department of Radiology, e.g., x-rays of bone fractures sent to ER (the Emergency Room). In many cases, a radiographer will produce this material without involving the radiologist; however, the radiologist may be involved in certain situations.

- Radiology reports based on analysis of radiology imagery, that goes into the patient record. Radiology reports are the department's most important delivery since it forms the basis for a diagnosis of the patient.

During the interview, it was discussed whether a machine learning algorithm could be a possibility for analyzing x-rays of the thorax. The interviewees said that in order to diagnose a thorax problem, the data material would have to be a combination between looking at the pictures and understanding who the patient is: "*It is clear that if you have to analyze the picture, then it is not only a question looking at the pictures but also to gather the story on old and new pictures.*" (15:26)

Additionally, it was said that this information is needed to understand what diagnosis or other analyses to look for: "*... information about the patient point directly towards what to look for*" (16:28), and "*... this information is used to be able to suggest further analyses.*" (16.39)

The complexity of analyzing the data material for one patient increases significantly if further analyses are taken account of: "*... Then you can go over to look at CT scannings, cross-section image - there you suddenly have significantly much bigger data materials and an infinity of things you potentially can look for.*" (16.54)

However, the large data material is in itself not necessarily just good. When the talk was about screening with a lot of data the interviewees said: "*.. if you take a blood sample there is 95 percent confidence interval for a normal test, so if you take 20 samples, there will be one error, if you take 100, there will be five errors - in that way you create more noise by screening generally.*" (22:50)

And it was added: "*... and more screening tests can create noise and be potentially harmful to the patient*" (psychological or physical side effects of a more invasive analysis following the blood sample). (23:23)

The interviewees referred to the Stanford breast cancer trial, mentioned in the previous section: "*... there was a test with computer-aided detection of breast cancer at Stanford. I will claim that that algorithm in half a day could be trained to look at the thorax and would identify 100 points to follow up on. And then next day, like a computer, you can pick perhaps 5 or 10 spots that need more investigation*" (20:23), and "*... and* *then they are being directed towards a CT scanner, and that capacity is perhaps missing.*" (20:49)

The interviewees here identify a conflict concerning the amount of data that can be taken into consideration from a patient, and how that can become a problem if there is automatic screening of x-rays when there are no resources for bringing all the "maybe" identified patients in for further investigation and analysis.

Throughout the interview, it was discussed whether machine learning could be applied to x-rays of the thorax. The interviewees did not reject it completely but had different viewpoints and challenges they could see: "... *I think the approach that it is a decision-supporting system to improve quality is the right way to go, and I think we as radiologists should take ownership of that in the processes we have here and now.*" (20:07)

They identified concerns in a number of areas "... *If you take such an algorithm and use it on Danes, and then bring it to another place, my immediate guess would be that it would fail because there suddenly are new and other illness categories and another patient group.*" (18:50)

They agreed that it would be beneficial if it was possible to define when x-rays are normal to avoid a heavy workload: "... *One place where this could help would be to select what is normal. We spend really really a lot of hours looking at normal x-rays of the thorax*" (21:45), and "... *If you can find a limit for when it is 100% normal it would be good*"} (21:55), and furthermore " ... *so if we could get a filter to start with, that would be one way of doing it.*" (22:30)

Radiology reports are, as with most of the patient record, written in freestyle prose. This format means that the information contained in the patient records are mostly unstructured data. There is no apparent reason why radiology reports do not contain some information as structured data, but the reports have historically been produced this way.

However, some parts of the patient record are strictly structured and do not contain any diagnostic analysis, as is the case for test results of blood samples. These are simply listed as measured values along with expected value ranges. This also means that the radiologists need to base their diagnoses on a holistic evaluation of everything: "... *We analyze the pictures for classical things and some things which are less clear. The conclusion is based on the interpretation of the radiology department as well as the description, not only the pictures*" (43:30), and "... *It is typically an interpretation of what is there in the system. If there are consequences, it should be part of the description.*" (44.16)

The primary responsibility of the expert radiologist is to analyze radiology imagery. This material comes from several sources, such as x-ray, magnetic resonance imaging, computed tomography, among other sources, in 2D or 3D format. As it was mentioned in the interview: "... *varies the resolution to get a cut through many pictures to get more detail. That happens automatically with pictures that are made through CT scanning*"} (35.17), and "*.. one of the problems as radiologists is that our sensitivity, in theory, is rather high. You can see many things when scanning. Our specificity is not as good that these algorithms do not solve that. It raises our confidence in what we can see but not understanding what it is..*" (40.20), and furthered "... *The more things we see, the more things we need to relate to*" (40:50) "... *The systems can raise the sensitivity but are missing the specificity (the classification and confidence).*" (41.31)

The computer system used for analysis allows the radiologist for panning and zoom, and in some cases to apply filters to the imagery. However, the analysis is entirely based on the training and experience of the radiologist, in that the computer system does not assist with any automated augmentation or analysis.

The radiologist has access to the patient record, so medical history, age, and other relevant information can support the radiologist in the analysis. The patient record is an essential element, especially with thorax x-rays, as this kind of radiology imagery is difficult to analyze. Various pathologies can look very similar on an x-ray, so knowing that one is dealing with an elderly patient with several smoking-related ailments could suggest cancer rather than pneumonia.

With x-ray imagery, which is limited to 2D, the radiologist will usually know the diagnosis within a one-second time frame, as mentioned by one of the interviewees: "*I think you have made your decision about what you see in such a picture within one second.*" (1:21:10)

There is an actual risk of over-analyzing radiology imagery, mistaking noise and irregularities for pathologies. Over-analyzing could lead to more invasive medical examinations, exposing the patient to unnecessary risk of complications and pain. Thus, it is equally essential to exclude potential pathologies in the diagnosis, as to include pathologies. Other concerns are unnecessary costs to the healthcare system or patient and limited capacity of other advanced medical equipment. As expressed in the interview "... *maybe a system will create a possibility to look at data in other ways than we do today. There are 1000 other ways to analyze the data than what we do, and much qualitative data we do not look at today.*" (1:03:50)

An additional idea of what a machine learning algorithm could do was to identify spots that need to be looked at on the x-rays: "... *it would be relatively easy to do a program that made arrows on the pictures so it would be relatively easy for us to look at them afterward. We do not have that at the moment.*" (37.20)

They can see the possibility for bringing new confidence and support to the diagnostic setting: "... *I would wish for a system that could increase my certainty in what I see in the pictures. If it says it is 80% plus then I will be more certain*"} (1:05:05), "... *Personally, I would not mind if the system could provide five diagnosis suggestions with the highest likelihood attached to them*" (1:05:56), and "... *And if you had computer systems that could calculate the volume of tumors! That would be fantastic*" (1:13:15)

The interview ended by the interviewees stating that they themselves are fond of computers and interested but that a lot of other users not think in the same way and that therefore this can create some obstacles for implementing such systems.

V. Analysis

In our case study interview, we identified two themes that kept coming up. The first theme was the importance of both being able to detect and reject pathologies. From a technical standpoint this means that proper prioritization of sensitivity and specificity is of utmost importance. The second theme was the context in which the diagnosis is given. The radiologist will seek out information in the patient record to support his diagnosis and make him more confident in his decision, but the process is very reliant on the experience and expertise of the radiologist. However, other data than the data in the radiology imagery is essential to the radiologist's performance.

The most successful applications of artificial intelligence and machine learning outside academia so far have been applications of supervised learning, or more specifically, solving classification and regression problems based on a large amount of labeled data.[9] Furthermore, the problems that we are most likely to solve successfully with machine learning are problems that can be solved in a one-second time frame by a human being.[9] As discussed in section IV, this is precisely the case when the radiologist examines radiology imagery such as x-rays, and thus this kind of work should be well suited for automation.

Sokolova et al. have found that research in machine learning in most cases report performance of machine learning algorithms using basic metrics such as accuracy, precision and sensitivity, F(1) score, and Receiver Operating Characteristic (ROC).[19] Sokolova et al. also demonstrated how these performance metrics can be misleading to the degree of coming to the wrong conclusion, and should be avoided in cases where classes of classification are equally important, and data is expensive to generate.[19] While there might be an abundance of data available in healthcare, the process of preparing and labeling data for use in a supervised learning setting is both labor intensive and expensive, because hundreds of thousands of samples are needed, if not more.[9]

During the interview, it is mentioned that one problem with screening data, and generally about having an abundance of data is that it may generate "noise" or false positives which is a problem for both the hospital as well as for the patient. In medical research, identifying true positives and avoiding false positives is of equal importance,[20] as discussed in section IV. To address this issue, Youden proposed the Youden index in 1950 [20] that “*equally weights the algorithm’s performance on positive and negative examples.*”[19] Other metrics with similar properties exist, e.g., the likelihood and discriminant power metrics, and all these metrics are already widely used in medical research.[19] With this in mind, it would be useful and interesting to know how Rajpurkar et al.,[4] Wang et al.,[11] and Yao et al. [12] would perform using these metrics, and how the expert radiologists from the Rajpurkar et al.,[4] study performed.

The other area that was of most concern during the interview was the decision on the diagnosis of the patient and the complexity of the process for deciding on that. As we learned from our case study in section IV, information from the patient record is essential in the analysis of radiology imagery, especially in the case of x-ray imagery. The patient record and the newly obtained radiology imagery combined form a *diagnostic context* that is the basis for the radiologist's diagnosis. Therefore, while valuable and remarkable from a technical perspective, the comparison of diagnostic performance between the machine learning system and the expert radiologists done by Rajpurkar et al. [4] is not directly applicable in a clinical setting. Even if no patient record is available, e.g., if dealing with a patient of unknown identity, it would still be possible to make basic observations about the patient such as age and general condition.

The interviewees stated that they could not currently see the application of a machine learning algorithm as a diagnosis system, but they see possibilities for applying machine learning in some steps of the diagnosis process. One of the statements from the interviewees is that the possibility for such a system actually pointing to a high percentage of the normal thorax x-ray pictures would be tremendously helpful and would save significant amounts of time (they mention that they look at a lot of normal thorax x-ray pictures). Another comment from the interview, suggests that a decision support system could aid the diagnosis process by having pointing arrows on the x-rays to show, which parts of the x-ray have a higher likelihood of being abnormal and therefore something that needs extra attention.

VI. Conclusion

We have analyzed the possibility of applying machine learning-based radiology diagnosis in a clinical setting. We have focused on the analysis of thorax imagery, and drawn from prior research done by Rajpurkar et al. [4]

We find that while machine learning could be applicable as a decision-supporting tool, it is unlikely that, by itself, it could perform actual diagnosis in a clinical setting. This is due to the fact that the diagnostic process is extremely complicated, and depends on structured and unstructured data from several sources in the patient record. While machine learning systems are able to generalize diagnoses from imagery, the diagnostic process consists of several other aspects than the imagery itself, e.g., patient history, and observations from other examinations.

However, we found that machine learning systems would be useful in radiology for various other tasks, e.g., augmentation of imagery, classification of high-confidence non-pathological patients (that would not need further analysis).

Finally, we found a need for alignment of the reporting of machine learning research performance with established metrics from medical research. We suggest further study in fitting machine learning metrics to examples of real-world environments would be an important stage in moving machine learning from research environments into daily professional practice.

Acknowledgment

We would like to acknowledge Chief Physician Martin Lundsgaard Hansen, MD, and Chief Physician Thomas Axelsen, MD, at Department of Radiology, Rigshospitalet in Denmark, whose help and expertise made this study possible.

We would also like to acknowledge Jørn R. Strøm, MD, for providing insights into the Danish healthcare sector during many conversations.

References

[1] "MACHINE LEARNING: BLIV KLOGERE PÅ TEKNOLOGIEN BAG KUNSTIG INTELLIGENS," *IDA Universe*. [Online]. Available: https://universe.ida.dk/tema/machine-learning/. [Accessed: 15-Nov-2017].

[2] M. Strumpe and M. Craig, "Applying Deep Learning to Metastatic Breast Cancer Detection," *Google AI Blog*, 2018. [Online]. Available: https://ai.googleblog.com/2018/10/applying-deep-learning-to-metastatic.html. [Accessed: 15-Nov-2018].

[3] E. Corbrett, "The Real-World Benefits of Machine Learning in Healthcare," *HealthCatalyst*. [Online]. Available: https://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare.

[4] P. et al. Rajpurkar, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv Prepr. arXiv1711.05225*, p. 7, 2017.

[5] Y. Engestrom, "Activity theory as a framework for analyzing and redesigning work," *Ergonomics*, vol. 43, no. 7, pp. 960–974, 2000.

[6] P. R. Carlile, "Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries," *Organ. Sci.*, vol. 15, no. 5, pp. 555–568, 2004.

[7] T. Greenhalgh and D. Swinglehurst, "Studying technology use as social practice: The untapped potential of ethnography," *BMC Med.*, 2011.

[8] S. Jones, N. A. M. Maiden, S. Manning, and J. Greenwood, "Informing the specification of a large-scale socio-technical system with models of human activity," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*, 2007, pp. 175–189.

[9] A. Ng, "What artificial intelligence can and can't do right now," *Harv. Bus. Rev.*, vol. 9, 2016.

[10] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993.

[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017, pp. 3462–3471.

[12] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv Prepr. arXiv1710.10501*, 2017.

[13] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[14] A. Ng, "Should radiologists be worried about their jobs?," *Twitter*, 2017. [Online]. Available: https://twitter.com/AndrewYNg/status/930938692310482944. [Accessed: 13-Sep-2018].

[15] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[16] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," *J. Biomed. Inform.*, vol. 42, no. 5, pp. 760–772, 2009.

[17] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, "Natural language processing in radiology: a systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[18] C. Osborne, B. Frey, and M. A, "The future of employment: how susceptible are jobs to computerisation?," *Technol. Forecast. Soc. Change*, vol. 114, pp. 254–280, 2017.

[19] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation," *Adv. Artif. Intell.*, vol. 4304, pp. 1015–1021, 2006.

[20] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.