Requirement analysis for an artificial intelligence model for the diagnosis of the COVID-19 from chest X-ray data

Tuomo Kalliokoski
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
tuomo.kalliokoski@jyu.fi

***Abstract*—There are multiple papers published about different AI models for the COVID-19 diagnosis with promising results. Unfortunately according to the reviews many of the papers do not reach the level of sophistication needed for a clinically usable model. In this paper I go through multiple review papers, guidelines, and other relevant material in order to generate more comprehensive requirements for the future papers proposing a AI based diagnosis of the COVID-19 from chest X-ray data (CXR). Main findings are that a clinically usable AI needs to have an extremely good documentation, comprehensive statistical analysis of the possible biases and performance, and an explainability module.**

***Index Terms*—COVID-19, AI, CXR, requirement analysis**

I. Introduction

Ever since the World Health Organization classified the COVID-19 as a Public Health Emergency of International Concern (PHEIC) [1], which is more commonly called as a pandemic [2], the AI field has produced a multitude of papers related to diagnosing the COVID from various data. Reviews which focus on the clinical suitability of the models presented like references [3], [4], [5], [6], [7], and [8] have been critical about various aspects of the reviewed papers. Thus there is a need for a proper requirements analysis for the AI based diagnosis of the COVID-19 from CXR data so that these shortcomings can be remedied.

This requirement analysis includes general ethical considerations, general AI model building considerations, and clinical considerations for both the AI in general and in the radiology diagnosis. For these considerations I will use relevant publications as sources of the requirements. Many of the other sources are reviews of previous work in which I will concentrate on the criticism of the analyzed publications.

In the next section I will go through the selected sources in order to find the information about requirements. This is followed by a section in which I formulate those requirements in a more concrete way. The fourth section is the proposed solutions for the given requirements. The last section is the conclusions.

II. Search for the requirements

A. General ethical considerations

The issue of the ethics in the AI and data field is a field of study in its own right [9] as is the field of the medical ethics. I will refrain from the full discussion of these matters and focus on the issues from the general computer science ethics point of view. There is plenty of literature available about the ethics for AI in healthcare see for example references [10], [11], [12], [13], [14], [15], and [16].

In the software development practically every design decision has to be justifiable after ethical analysis [17]. Simplest question is about how resource intensive can we be, how much do we value the output accuracy versus the resource usage to get that accuracy? Each technological project should also undergo proper impact assessment [18].

This analysis is based on the framework provided in the reference [18] which provides a list of questions for consideration. I will go through most of them in order, but skip those which are trivially irrelevant to this work.

Questions about respect for the autonomy are mostly trivially irrelevant, except the case of curtailing personal freedom of movement. If a person is diagnosed with the COVID he will most likely be quarantined. This is justified due to the public health risk which the infected persons pose to the others.

In the area of the dignity the questions are mostly trivially irrelevant, as the goal is to provide a quicker and less invasive way of diagnosing the COVID from patients with a pneumonia. The patients would be in any case subjected to the chest X-rays and the diagnosis tool would be used to analyze those images.

Next section in the reference [18] is about the informed consent. In the case of the data collection one should be using datasets only from respectable institutions which have followed the required ethical practices in their work. When the AI model would be used by the medical practitioners can only give their informed consent if the quality of the AI system is evaluated properly and it has an explainability functionality. This explainability is required by the medical practitioners to do informed decisions based on the output of the system [19].

Another issue here is the collection on data during the use of the system.

Non-maleficence section of the reference [18] starts with safety related questions. As this would be used as a clinical tool the safety aspect is a critical one. There will be a great harm coming from the errors in the diagnosis. A false positive diagnosis will cause psychological harm to the patient and a false negative will slow down the treatment and puts other people in risk as well. The algorithm needs a thorough testing before it can be used in a clinical settings.

Second part of this section in the reference [18] is the social solidarity, and inclusion and exclusion. This is mainly about the information society inclusion and only relevant to this discussion is the fact that the system should be available for offline use in the areas where the internet connectivity is not good.

The beneficence section of the reference [18] has multiple questions which are relevant to this work. The goal is to benefit the individuals and the society by a faster and less invasive way to diagnose the COVID. With the X-ray image analysis the diagnosis can be done in 30 minutes [20], while the nasal swap and the RT-PCR will take at minimum multiple hours to complete [21]. This should be a great benefit for all humans.

Next section in the reference [18] is the universal service. It should be available in all medical stations equipped with the X-ray machinery and a computer. Here one should also take into account the issue of the global computer part shortage which limits the ability to get the newer and more powerful computing equipment [22], [23], [24], [25]. The diagnosis software thus should be usable in any modern computer.

The accessibility is also an issue which needs some discussion. In the case of fully usable software an extremely simple user interface so that it is easy to use with a minimal training. If the goal is only produce the model for diagnostic, it has to have a simple and well documented API.

The value sensitive design has some relevance here. The explainable nature of the AI will provide empowerment to the medical personnel. With a "regular" AI they will get only a value stating that the patient has the COVID with some probability, with an eXplainable AI they also get information on why the AI has come to this conclusion. This will provide them a lot more information and they can use it for their benefit.

For the sustainability we have some issues with possible change in the standards. The system should be built using the current standards and a modular design so that it could be easily updated.

The justice section in the reference [18] discusses about distributive justice for all individuals and groups. This is highly relevant matter for any diagnosis tool as it is widely known that there are problems with many diagnosis methods when the patient is not a Caucasian male. See for example references [26], and [27]. To overcome this problem the dataset needs to have data with an excellent reach over all humanity, not just in a single demographic group. The less desirable alternative solution is to state clearly the issues and limitations of the diagnosis model in the publication.

The equality and fairness (social justice) as defined in the reference [18] is not as relevant as the previous part. The main point in it is the availability of the service for all, not just to a segment of the population based on their privileges. This is solved by the same solution as the universal service issues. Another point raised is the risk for the diagnosis being used for detriment of the patient.

Next part in the reference [18] talks about privacy and data protection. This is mainly relevant to the training data which I have discussed earlier. These issues is addressed when selecting the data source and by not collecting data during the use of the system.

B. General Data Science and AI related considerations

Almost every AI project can be seen as a data science project in which we are tasked to find new knowledge from the data. For this there are well established workflows like KDD [28] and CRISP-DM [29], and others for which one can see reference [30] which provides a review.

Both the KDD and the CRISP-DM start with the understanding of the domain. One has to have enough domain knowledge to know what is needed and which things are relevant. This includes the knowledge on what has to be reported.

Next step is the data understanding and preparation. One has to know the data properly for identifying possible biases, confounding factors which could lead to shortcut learning [31], imbalance [32] and other possible issues in the data quality and suitability for the task. Data selection is one of the most critical tasks.

This is followed by the modeling. This requires selection of suitable modeling tools for the data and the goal.

The statistical model analysis is then performed, which includes proper review of models properties related to the goal. Here we need to remember that any data which has been used in the model building cannot be used in the evaluation [33], [34] (including those which are not independent from the data used in the building.) One should remember that there are multiple ways of doing the estimation of the generalization performance and choose the most appropriate one [35].

Last step is deployment in to the production.

The usual structure of the project is also iterative one, so based on the discoveries in each step one goes back to a suitable step and acts upon the changed situation. Like if you find confounding factors then you go to the data preparation to remove them.

For AI tasked to image analysis CNN [36], [37] is the default choice of the architecture, but there are options like Capsule Neural Network (CapsNet) [38], Graph Neural Network (GNN) [39], and their combination CapsGNN [40]. These all require proper understanding of the data for the proper architecture implementation. The decision between different options depends on the task specific details and data availability.

There are options for handling data imbalance [41], [42] and scarcity. Before using any of the standard image augmentation techniques [43] one must have enough domain knowledge to refrain generating incorrect data, as a simple example I use rotating 6 to 9 in number recognition. Example of this is found in the figure 1.

C. The analysis of the reviews done to the previous COVID-19 diagnosis models

There has been many papers published for reviewing work done to diagnose the COVID-19 from the CXR- and CT-images. Reviews presented in this section are focused on the clinical applicability of the reviewed models and thus are extremely critical as the field has extremely strict regulation and tight tolerances. I have used these reviews for gathering requirements.

The reference [6] was among the first ones and it studied 14 publications of the AI models for the COVID-19 detection published by the end of March 2020. It found issues with testing for biases, the used datasets, and with used algorithms, including the use of out-of-the box models and lack of the explainability.

The use of different classification methods (binary, multiple classes, multiple labels, and hierarchical) in the COVID diagnosis was researched in the reference [7]. They found 11 studies related to their interest by May 5 2020. They found out that on those there were lack of consistency in the evaluating the quality of their model predictions.

Publications between March 2020 and May 2020 were also reviewed in the reference [8]. They found 34 publications and found that many publications had issues with dataset selection and possibly used multiple copies of same images. Other reported issues with used data were the class imbalances, private datasets. They also criticized the lack of uniformity in the quality evaluation, including bias evaluation. Explainability was also an issue which they brought up.

According to reference [4] studied papers proposing ML based diagnosis of the COVID-19 with data from the CXR or the CT scans from 1 January 2020 to 3 October 2020. They found 320 papers for their quality review and 258 failed in the first section of their analysis. Insufficient documentation in the model selection (132 failures), the methods of preprocessing of the images (125), and the details of the training approach (105) were the three most common failures. One critical failure was not disclosing the dataset used in the analysis. For the papers which passed this screening among the reasons for failing the clinical suitability were the lack of the proper validation, the robustness or sensitivity analysis, the demographics of the people in the data, the statistical testing for results, and the reporting issues regarding to the generalization. The paper criticized the lack of attention given to the features of the used datasets, for example using the dataset [44] as the control set while it consists of pediatric patients aged between on and five, and the COVID-19 patients were adults. Another issue raised was the downward scaling of the images due to the use of the ready of-the-shelf models. This and the lack of the demographic data is also related to the use of JPEG and PNG images instead of the DICOM [45] which has metadata of the image acquisition parameters and other important information. The code availability and other replicability issues were also mentioned as was the call for the interpretability.

Another paper dedicated for creating proper basis for responsible deep learning for diagnosing COVID-19 from medical images is reference [3]. This paper is centered around use of the Explainable Artificial Intelligence to find the possible errors in the model, but was not limited to it. They analyzed 25 models, which were collected by August 14 2020. They found mistakes in data acquisition, model development, and explainability. As an example the paper mentioned that one should not use image augmentations which produce ”impossible” images. For future models it produced a checklist for creating a responsible deep learning model for this task.

Reference [5] presents a systemic review of 169 studies. It provides a critical appraisal of the prediction models for the diagnosis and the prognosis of the COVID-19 in the selected studies which also included other implementations than the image analysis of CXR or CT. They found that all studies had either a high or an unknown risk of a bias in the results. This was mainly caused by non-representative selection of control patients, overfitting, issues with result validation, and unclear reporting.

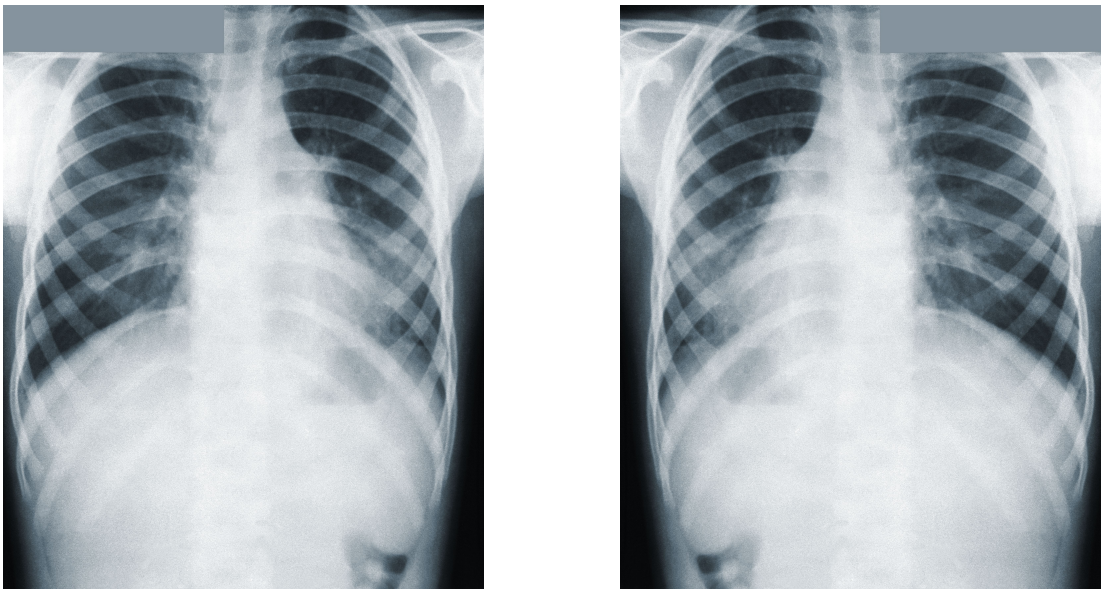
Many other papers have raised the issue of lack of generalization of the AI diagnosis due to confounding factors in the data [46], [47], [48], [49], [50], [51]. This leads to diagnosis based on other issues than the actual relevant features on the lung area, like learning from which data set the image is based on other information visible in the image. Another recognized source for lack of generalization was the dataset biases [52], [53], [54], [51]. Possible sources for these biases include patient demographics, procedures (for example the direction from which the X-ray image was taken), and procedures performed before taking the X-ray (for example an intubation tube visible in the CXR image.)

Reference [55] calls for uncertainty evaluation for predictions. This is related to lack of statistical analysis of the prediction quality mentioned in both references [3] and [4].

D. Analysis of other AI related medical publications

There are papers written on the use of the AI for diagnosis or for other purposes in the medical science. In this section I will discuss the issues which they have raised.

The medical papers on the AI use outside of the COVID-19 are also relevant as they provide information on what is needed from a AI product used in the medicine. Reference [56] called for the good training data, the performance validation, and was critical of the ”black-box” nature of some AI models. Papers like [57], [58], [59], [60], and [61] reminded on issues with the data and the ”black-box”. Reference [61] even called the ”black-box” AI as unacceptable in the medical domain. The lack of the documentation for reproducibility was also brought up by the reference [60]. Among other papers calling for the



(a) Original CXR (b) Flipped CXR

Fig. 1: An example of incorrect augmentation, note how the internal organs are in incorrect position after flipping the image horizontally. Original photo is Public Domain from CDC.

XAI are references [62], [63], [64], [65], [66], and [67]. The reference [68] points out issues with amount of data available for training and imbalance issues in the data, and it also point out the lack of the confidence intervals in the predictions.

The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) is available in the reference [69]. There is also a radiomics quality score (RQS) [70], which can be used.

Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [71] gives a reporting guidelines for diagnosis tools.

The PROBAST (Prediction model Risk Of Bias ASsessment Tool) [72] tool is commonly used for the estimation of the risk of bias. The bias is usually related to the dataset not being representative due to the data scarcity, the population shift, the prevalence shift, or the selection bias [73], [74]. The dataset shifts are also discussed in reference [75].

The reference [76] noted confounding factors in the medical image analysis of the knee is also.

The reference [77] describes the application of the ITU/WHO FG-AI4H (Focus Group on Artificial Intelligence for Health) assessment guidelines [78] for a machine learning tool. The application includes a questionnaire about the project, the bias & fairness analysis, the interpretability (explainability), the robustness evaluation and the reporting guidelines.

III. Requirements found

A. Ethical requirements

In this section I will transfer the thoughts presented in the ethical considerations section into more proper requirements.

1) The data collection has to be done with an informed consent or use data from a collection by a respectable source.
2) The quality of the diagnosis has to be known to the medical practitioners.
3) Medical practitioners needs to have explanation why the software made the diagnosis.
4) Product needs to be usable in a low infrastructure area. This means that the software should be able to be used without an internet connection and with a low end computer.
5) Faster or better than the competing technologies (see for example the references [20], [21], and [79]). The diagnosis quality better than rapid antigen testing and result should be available in less than 10 minutes with an standard medium or low power computer.
6) Good and simple user interface. It can be limited to a simple and well documented API.
7) Use eXplainable AI to empower the users.
8) Use medical imaging standards (DICOM [45]).
9) The model should work as well with all segments of the human population. If this is not possible it should be clearly stated.
10) Do not collect data during the operation without consent.

B. General data science and AI requirements

Here we basically have only two main requirements.

1) Learn the domain issues.
2) Study the data properly to find any possible issues.

while others are

3) Select only proper data.

4) Fix any issues with the data.
5) Select suitable modeling tools for the data.
6) Perform proper analysis of the model created.
7) Repeat if result is not good enough.

C. Clinical AI requirements for the COVID diagnosis

In order to generalize the results (or at least know the limits of the generalization) we need to:

1) Remove the confounding factors.
2) Handle the biases in the datasets.
3) Handle the data scarcity.
4) Handle the shifts between the datasets and the reality.

For reliability review which is needed for every medical application we need to:

5) Do proper the documentation of choices made.
6) Do the result validation.
7) Do the robustness and sensitivity analysis.
8) Select the image size for the analysis based on scientific reasons, not based on the convenience.
9) Use the metadata available.
10) Have the code and other material available for replication.
11) Explain the reasons for given diagnosis.

D. Requirements from the other medical AI publications

Here we have following:

1) Use checklist(s).
2) Explainability.
3) Bias analysis.
4) Proper data issue handling.
5) Documentation.
6) Performance validation.

E. Combined requirements

Many of these requirements are overlapping and thus the list of the requirements can be simplified to following.

- Learn the basics of the medical imaging domain.
- Use checklist(s).
- Use the DICOM data from a respectable source.
- Study the available data properly.
- Handle data issues: selection, biases, confounding factors, scarcity, and shifts.
- Select a suitable model with explainability.
- Document every decision, and the reasons for them, regarding to the data, the model architecture, and the (meta)parameters of the model.
- Do the proper statistical analysis of the quality of the model predictions.
- The model needs to be better than other diagnosis methods. (Faster, more accurate, or better availability in low infrastructure areas.)
- Store everything needed for the replication of the results.

IV. Practical solutions to these requirements

The requirements and solutions listed here are aimed to be used for the preliminary studies to find suitable models. For the proper handling the issues do take a look at the references given here and previously in this work and legal requirements for clinical applications. My recommendations are also found in a condensed form in the table I.

A. Learn the basics of the medical imaging domain.

The best solution here is to include a domain expert into the group, the absolute minimum is a proper review of the previous work done in the field.

B. Use checklists

There are multiple checklists e.g references [3], [69], [72], [78], [70], [71]

C. Use full sized DICOM data from a respectable source

This has its own solution written directly in to the requirement. As DICOM is an industry standard it is well documented and there exists ready libraries for its use. The respectable source would be some proper institute, not a private collection.

D. Study the available data properly.

Do proper analysis to the raw data. This has to include study of demographics, biases, duplicates, outliers, pixel intensities, etc.

E. Handle data issues: Selection, biases, confounding factors, scarcity, and shifts.

In the data selection one needs to be careful and one has to remove the confounding factors and the incorrect data. Especially the area outside of lungs holds many confounding artifacts [3], while the secondary source of confounding is the pixel intensity [50]. There are plenty of ready lung segmentation models available, but they should be reviewed properly before selection. Incorrect data should not be included, examples of this are duplicates and failed images.

In the case of the bias this can sometimes be handled with the proper selection of the used data or with generating new data via the augmentation [80], [81] or via the generative adversarial nets (GAN) [82]. The last resort is to record properly the existing biases and continue with them. On the new data generation one needs to be extremely careful not to use incorrect techniques [3].

For data scarcity we have a possibility to generate new data as in the case of bias and other solution is the transfer learning [83], [84], [85], [86]. But one should remember that transfer learning is not always useful [87], [88].

Dataset shifts are discussed with detail in the references [73] and [75].

F. Select a suitable model with explainability.

CNN is still the default choice, but if other relevant factors points toward choosing something else do not discard them. The available out of the box models are not always the best choice [87], [81].

There are multiple different explanation tools for the image classification AI. The reference [3] gives some overview what has been used previously and gives some pointers on the issues related to them. The suitability of these explanation tools for other than the CNN is also an issue.

G. Documentation

Document every decision, and the reasons for them, regarding to the data, the model architecture, and the (meta)parameters of the model.

H. Do a proper statistical analysis of the quality of the model predictions.

According to the reference [7] we need following statistical information.

Binary categorization
: Accuracy, precision, recall (sensitivity), F score, specificity, AUC.

Multi-class classification
: Average accuracy, error rate, precision$_\mu$, recall$_\mu$, F score$_\mu$, precision$_M$, recall$_M$, F score$_M$

Multi-label classification
: Exact match ratio, labeling F score, retrieval F score, Hamming loss.

Hierarchical classification
: Precision↓, recall↓, F score↓, precision↑, recall↑ and F score↑, which are defined in the reference [7].

I. Model performance

Compare the model to the current state of the art with other techniques like the RT-PCR. Remember to use up to date information on this comparison.

J. Store everything needed for replication of the results

Use the GitHub or some other similar service.

V. Conclusions

There has been a great effort and lots of enthusiasm for providing an AI solution to the clinical diagnosis of the COVID using the CXR data. While promising results have been published, unfortunately most of the publications lack the rigor needed in the medical field. This issue is shown by multiple reviews [3], [4], [5], [6], [7], and [8] and our field needs to pay a proper respect to the actual requirements for such tools.

This work is a start into this direction and a pointer towards more thorough work done by the domain experts.

References

[1] World Health Organization, "Novel coronavirus (2019-nCoV): situation report, 11," World Health Organization, Technical documents, 2020-01-31. [Online]. Available: https://apps.who.int/iris/handle/10665/330776

[2] ——, "Coronavirus disease 2019 (COVID-19): situation report, 51," World Health Organization, Technical documents, 2020-03-11. [Online]. Available: https://apps.who.int/iris/handle/10665/331475

[3] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, and P. Biecek, "Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies," *Pattern Recognition*, vol. 118, p. 108035, 2021.

[4] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[5] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[6] A. Burlacu, R. Crisan-Dabija, I. V. Popa, B. Artene, V. Birzu, M. Pricop *et al.*, "Curbing the AI-induced enthusiasm in diagnosing COVID-19 on chest X-rays: the present and the near-future," *medRxiv*, 2020.

[7] O. S. Albahri, A. A. Zaidan, A. S. Albahri, B. B. Zaidan, K. H. Abdulkareem, Z. T. Al-qaysi *et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *Journal of Infection and Public Health*, vol. 13, no. 10, pp. 1381–1396, 2020.

[8] H. S. Alghamdi, G. Amoudi, S. Elhag, K. Saeedi, and J. Nasser, "Deep learning approaches for detecting COVID-19 from chest X-ray images: A survey," *IEEE Access*, vol. 9, pp. 20 235–20 254, 2021.

[9] E. Kazim and A. S. Koshiyama, "A high-level overview of AI ethics," *Patterns*, vol. 2, no. 9, p. 100314, 2021.

[10] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *The American Journal of Bioethics*, vol. 20, no. 11, pp. 7–17, 2020, pMID: 33103967.

[11] J. R. Geis, A. P. Brady, C. C. Wu, J. Spencer, E. Ranschaert, J. L. Jaremko *et al.*, "Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement," *Radiology*, vol. 293, no. 2, pp. 436–440, 2019, pMID: 31573399.

[12] F. Pesapane, C. Volonté, M. Codari, and F. Sardanelli, "Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States," *Insights into Imaging*, vol. 9, no. 5, pp. 745–753, 2018. [Online]. Available: https://doi.org/10.1007/s13244-018-0645-y

[13] I. de Miguel, B. Sanz, and G. Lazcoz, "Machine learning in the EU health care context: exploring the ethical, legal and social issues," *Information, Communication & Society*, vol. 23, no. 8, pp. 1139–1153, 2020.

[14] M. Mirbabaie, L. Hofeditz, N. R. J. Frick, and S. Stieglitz, "Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research," *AI & SOCIETY*, 2021. [Online]. Available: https://doi.org/10.1007/s00146-021-01239-4

[15] J. Morley, C. Machado, C. Burr, J. Cowls, M. Taddeo, and L. Floridi, "The debate on the ethics of AI in health care: a reconstruction and critical review," *SSRN Electronic Journal*, 2019.

[16] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo *et al.*, "The ethics of AI in health care: A mapping review," *Social Science & Medicine*, vol. 260, p. 113172, 2020.

[17] F. Kraemer, K. van Overveld, and M. Peterson, "Is there an ethics of algorithms?" *Ethics and Information Technology*, vol. 13, no. 3, pp. 251–260, 2011.

[18] D. Wright, "A framework for the ethical impact assessment of information technology," *Ethics and Information Technology*, vol. 13, no. 3, pp. 199–226, 2011.

[19] A. I. F. Poon and J. J. Y. Sung, "Opening the black box of AI-medicine," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 3, pp. 581–584, 2021.

[20] L. Pan, J. Zeng, H. Pu, and S. Peng, "How to optimize the radiology protocol during the global COVID-19 epidemic: Keypoints from Sichuan provincial people's hospital," *Clinical Imaging*, vol. 69, pp. 324–327, 2021.

TABLE I: Requirements and proposed solutions

Requirement	Solution
Learn the domain	Get a domain expert or read the proper source papers.
Use checklists	Lists are found in references [3], [69], [70], [71], [72]
Data gathering	Use the DICOM data from a respectable sources
Data preview	Do a proper preview of the data properties to find the possible issues.
Data issues	Handle the issues with care, see at least the references [3], [73], and [75].
Select model	Find the proper model suitable for the task and equip it with the explanability module.
Documentation	Document and explain all choices made to the regarding data, the architecture, and the (meta)parameters.
Statistical analysis	Follow the reference [7]
Model performance	Compare to other technologies like the RT-PCR
Replication	Store everything needed for it in the GitHub or in an other similar service.

[21] N. Jawerth, "How is the COVID-19 virus detected using real time RT-PCR?" *IAEA Bulletin (Online)*, vol. 61, no. 2, pp. 8–11, 2020. [Online]. Available: https://www.iaea.org/sites/default/files/6120811.pdf

[22] M. Cooney, "Chip shortage will hit IT-hardware buyers for months to years: Tech executives and analysts say the current processor-chip shortage and disruption of supply chains thanks to COVID-19 could have a long-term impact on price and availability," *Network World (Online)*, 2021. [Online]. Available: https://www.networkworld.com/article/3619210/chip-shortage-will-hit-it-hardware-buyers-for-months-to-years.html

[23] The Economist Intelligence Unit, "The global chip shortage is here for some time: Loading, please wait," *Global Business Review*, 2021. [Online]. Available: https://www.economist.com/finance-and-economics/2021/05/20/the-global-chip-shortage-is-here-for-some-time

[24] A. Patrizio, "You're not imaging things, there is a serious chip shortage: CPUs, GPUs, and memory are all in tight supply due to manufacturing issues and high demand," *Network World (Online)*, 2021. [Online]. Available: https://www.networkworld.com/article/3623753/the-chip-shortage-is-real-but-driven-by-more-than-covid.html

[25] "Chips in a crisis," *Nature Electronics*, vol. 4, no. 5, pp. 317–317, 2021. [Online]. Available: https://doi.org/10.1038/s41928-021-00601-0

[26] K. Mahendraraj, K. Sidhu, C. S. M. Lau, G. J. McRoy, R. S. Chamberlain, and F. O. Smith, "Malignant melanoma in African-Americans: A population-based clinical outcomes study involving 1106 African-American patients from the surveillance, epidemiology, and end result (SEER) database (1988-2011)." *Medicine*, vol. 96, p. e6258, 2017.

[27] P. Silveyra and X. Tigno, Eds., *Sex-Based Differences in Lung Physiology*. Springer International Publishing, 2021.

[28] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery: An Overview*. USA: American Association for Artificial Intelligence, 1996, p. 1–34.

[29] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer *et al.*, "CRISP-DM 1.0 step-by-step data mining guide," The CRISP-DM consortium, Tech. Rep., August 2000. [Online]. Available: https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf

[30] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137–166, 06 2010.

[31] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[33] E. E. Cureton, "Validity, reliability, and baloney," *Educational and Psychological Measurement*, vol. 10, no. 1, pp. 94–96, 1950. [Online]. Available: https://doi.org/10.1177/001316445001000107

[34] A. K. Kurtz, "A research test of the Rorschach test," *Personnel Psychology*, vol. 1, no. 1, pp. 41–51, 1948. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.1948.tb01292.x

[35] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, vol. 2, no. 3, pp. 249–262, 2018. [Online]. Available: https://doi.org/10.1007/s41664-018-0068-2

[36] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980. [Online]. Available: https://doi.org/10.1007/BF00344251

[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3859–3869.

[39] M. Gori, G. Monfardini, and F. Scarselli, "A new model for earning in raph domains," vol. 2, 01 2005, pp. 729 – 734 vol. 2.

[40] Z. Xinyi and L. Chen, "Capsule graph neural network," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Byl8BnRcYm

[41] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. [Online]. Available: https://doi.org/10.1142/S0218001409007326

[42] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9004–9012.

[43] J. Shijie, W. Ping, J. Peiyi, and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," in *2017 Chinese Automation Congress (CAC)*, Oct 2017, pp. 4165–4170.

[44] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[45] *NEMA PS3 / ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard*, National Electrical Manufacturers Association Std., Rev. 2021d, 2021. [Online]. Available: http://www.dicomstandard.org

[46] G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from X-ray images," *Information Fusion*, vol. 76, pp. 1–7, 2021.

[47] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, pp. 1–17, 2018.

[48] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, 2021.

[49] K. B. Ahmed, G. M. Goldgof, R. Paul, D. B. Goldgof, and L. O. Hall, "Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification," *IEEE Access*, vol. 9, pp. 72 970–72 979, 2021.

[50] E. H. P. Pooch, P. Ballester, and R. C. Barros, "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification," in *Thoracic Image Analysis*, J. Petersen, R. San José Estépar, A. Schmidt-Richberg, S. Gerard, B. Lassen-Schmidt, C. Jacobs *et al.*, Eds. Cham: Springer International Publishing, 2020, pp. 74–83.

[51] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling COVID-19 from CHEST X-ray with deep

learning: A hurdles race with small data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, 2020.

[52] J. D. Janizek, G. Erion, A. J. DeGrave, and S.-I. Lee, "An adversarial approach for the robust classification of pneumonia from chest radiographs," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, ser. CHIL '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 69–79.

[53] S. Candemir and S. Antani, "A review on lung boundary detection in chest X-rays," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 4, pp. 563–576, 2019.

[54] C. Sáez, N. Romero, J. A. Conejero, and J. M. García-Gómez, "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset," *J Am Med Inform Assoc*, vol. 28, no. 2, pp. 360–364, 2021.

[55] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z. A. Sani *et al.*, "Objective evaluation of deep uncertainty predictions for COVID-19 detection," 2020.

[56] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[57] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E. Samir, O. S. Pianykh *et al.*, "Current applications and future impact of machine learning in radiology," *Radiology*, vol. 288, no. 2, pp. 318–328, 2018, pMID: 29944078.

[58] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[59] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo *et al.*, "Deep learning in medical imaging: General overview," *Korean J Radiol*, vol. 18, no. 4, pp. 570–584, 2017.

[60] K. Kallianos, J. Mongan, S. Antani, T. Henry, A. Taylor, J. Abuya *et al.*, "How far have we come? artificial intelligence for chest radiograph interpretation," *Clinical Radiology*, vol. 74, no. 5, pp. 338–345, 2019.

[61] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artificial Intelligence in Medicine*, vol. 106, p. 101878, 2020.

[62] J.-M. Fellous, G. Sapiro, A. Rossi, H. Mayberg, and M. Ferrante, "Explainable artificial intelligence for neuroscience: Behavioral neurostimulation," *Frontiers in Neuroscience*, vol. 13, p. 1346, 2019.

[63] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, no. 1, pp. 30–36, 2019.

[64] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017.

[65] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[66] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[67] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[68] F. Altaf, S. M. S. Islam, N. Akhtar, and N. K. Janjua, "Going deep in medical image analysis: Concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99 540–99 572, 2019.

[69] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e200029, 2020, pMID: 33937821. [Online]. Available: https://doi.org/10.1148/ryai.2020200029

[70] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, 2017. [Online]. Available: https://doi.org/10.1038/nrclinonc.2017.141

[71] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement." *BMJ (Clinical research ed.)*, vol. 350, no. jan07 4, p. g7594, 2015.

[72] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins *et al.*, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Ann Intern Med*, vol. 170, no. 1, pp. 51–58, 2019.

[73] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020.

[74] M. A. Al-masni, D.-H. Kim, and T.-S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Computer Methods and Programs in Biomedicine*, vol. 190, p. 105351, 2020.

[75] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," *Biostatistics*, vol. 21, no. 2, pp. 345–352, 2019.

[76] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale *et al.*, "Deep learning predicts hip fracture using confounding patient and healthcare variables," *npj Digital Medicine*, vol. 2, no. 1, p. 31, 2019.

[77] L. Oala, J. Fehr, L. Gilli, P. Balachandran, A. W. Leite, S. Calderon-Ramirez *et al.*, "ML4H auditing: From paper to practice," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, ser. Proceedings of Machine Learning Research, E. Alsentzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, S. Roy, and S. L. Hyland, Eds., vol. 136. PMLR, 2020, pp. 280–317. [Online]. Available: http://proceedings.mlr.press/v136/oala20a.html

[78] FG-AI4H, "The ITU/WHO focus group on artificial intelligence for health," 2021. [Online]. Available: https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx

[79] C. Leli, L. Di Matteo, F. Gotta, E. Cornaglia, D. Vay, I. Megna *et al.*, "Performance of a SARS-CoV-2 antigen rapid immunoassay in patients admitted to the emergency department," *International Journal of Infectious Diseases*, vol. 110, pp. 135–140, 2021.

[80] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[81] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of Healthcare Engineering*, vol. 2019, p. 4180949, 2019. [Online]. Available: https://doi.org/10.1155/2019/4180949

[82] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets

[83] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016. [Online]. Available: https://doi.org/10.1186/s40537-016-0043-6

[84] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds. Cham: Springer International Publishing, 2018, pp. 270–279.

[85] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "DeTrac: Transfer learning of class decomposed medical images in convolutional neural networks," *IEEE Access*, vol. 8, pp. 74 901–74 913, 2020.

[86] D. Karimi, S. K. Warfield, and A. Gholipour, "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations," *Artificial Intelligence in Medicine*, vol. 116, p. 102078, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365721000713

[87] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf

[88] V. Cheplygina, "Cats or CAT scans: Transfer learning from natural or medical image source data sets?" *Current Opinion in Biomedical Engineering*, vol. 9, pp. 21–27, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468451118300527