Accepted Manuscript of: Maria, F. et al. (2023). Trustworthy Applications of ML Algorithms in Medicine - Discussion and Preliminary Results for a Problem of Small Vessels Disease Diagnosis. In: Kowalczuk, Z. (eds) Intelligent and Safe Computer Systems in Control and Diagnostics. DPS 2022. Lecture Notes in Networks and Systems, vol 545. Springer, Cham. https://doi.org/10.1007/978-3-031-16159-9_1

Trustworthy applications of ML algorithms in medicine - discussion and preliminary results for a problem of Small Vessels Disease diagnosis

Ferlin Maria[1 [0000−0001−9286−0670]], Klawikowska Zuzanna[1 [0000−0001−6915−5041]], Niemierko Julia[2 []], Grzywińska Małgorzata[2 [0000−0001−6533−5774]], Kwasigroch Arkadiusz[1 [0000−0002−7803−0010]], Szurowska Edyta[2 [0000−0002−7042−4381]], and Grochowski Michał[1 [0000−0002−2453−2410]]

[1] Gdańsk University of Technology, Gdańsk, Poland
[2] Medical University of Gdansk, Gdansk, Poland

Abstract. ML algorithms are very effective tools for medical data analyzing, especially at image recognition. Although they cannot be considered as a stand-alone diagnostic tool, because it is a black-box, it can certainly be a medical support that minimize negative effect of human-factors. In high-risk domains, not only the correct diagnosis is important, but also the reasoning behind it. Therefore, it is important to focus on trustworthiness which is a concept that includes fairness, data security, ethics, privacy, and the ability to explain model decisions, either post-hoc or during the development. One of the interesting examples of a medical applications is automatic SVD diagnostics. A complete diagnosis of this disease requires a fusion of results for different lesions. This paper presents preliminary results related to the automatic recognition of SVD, more specifically the detection of CMB and WMH. The results achieved are presented in the context of trustworthy AI-based systems.

Keywords: Machine Learning · Artficial Intelligence · Deep Learning · Small Vessels Disease · Explainable AI · Trustworthiness.

1 Introduction

Machine learning algorithms have achieved a tremendous success in various image processing tasks. In particular, they obtain state-of-the-art performance in exploring and analyzing huge and complex datasets, especially images. Ones of the most important for society image recognition applications are the medical ones as an imaging is an integral part of medical diagnostics [1]. They include analysis of various type of image data, including 2D and 3D, from ultasonography (USG), computed tomography (CT), magnetic resonance imagining (MRI), endoscopy and others. In addition to careful analysis, a description of the examination and conclusions are needed. Medical data analysis is tedious and difficult. Moreover, the importance and sensitivity of this field makes it extremely important to describe it properly and clearly in order to make them trustworthy.

The clinician's description of the examination is biased in some ways. First of all, the analysis is subjective and based on the one's experience. Hence, sometimes it happens that two independent specialists assess the examination in different way, resulting in low observer reliability. Next, in some cases, there is a lack of standardized guidelines for evaluating the examination, while in others, despite many guidelines and rules, there are still some inconsistencies between descriptions provided by two specialists (e.g. for image data annotations). Another factor is the wide range of diagnostic equipment used, which varies depending on the manufacturer consequently influencing user's habits. Moreover, there are plenty of human-factors that strongly affect the evaluation process such as resting level, personal issues, and mood. Considering the above, the Clinical Decision Support System (CDSS) based on machine-learning [2]. seems to be a great opportunity to support clinicians in their daily work and minimize negative effect of human-factors while following guidelines. Such solutions can also support diagnostics in areas where access to specialists is limited. What is more, recent studies proved that the level of agreement of the AI tools with the experts was at least as good as agreement between two experts [3], and therefore, although the ML algorithms cannot be considered as a stand-alone tool, it can certainly be a valuable medical support.

In general, humans are reticent to adopt techniques that are not directly interpretable, tractable and trustworthy even if they reduce the bias of human participation. Deep learning (DL) models are such techniques. In high-risk applications like medical ones ceding medical decision-making to DL models without understanding of diagnosis rationale may violate the principle of non-maleficence and expose patients to harm. Therefore, in such applications, it is important to pay attention to system trustworthiness. From the clinician's point of view, in order to be acceptable, ML-based systems should have a level of transparency that allows their decisions to be verified by medical specialists in their level of domain knowledge. Explainable AI (XAI) techniques can provide effective tools to support this [4].

Small Vessel Disease (SVD) is a term which encompasses a variety of changes in human brain which are attributed to pathological changes in the small vessels. Anatomically, small vessels are arterioles, capillaries and venules. These structures are too small to be visible in CT or MRI. For this reason we focus on evaluating the lesions which appear as a result of pathologic changes to the small vessels. Among the factors which can lead to SVD we can distinguish: arteriolosclerosis, cerebral amyloid angiopathy, inherited/genetic small vessel diseases, CNS vasculitis, venous collagenosis or radiotherapy [5].

Basic tool for diagnosing SVD is neuroimaging, which makes it a task that can be automated by machine learning algorithms analyzing the image. Neurological changes can be visualized in both CT and MRI, with the latter being the current gold standard for diagnosis. For many years, there were no structured guidelines for reporting the imaging findings which was affecting the communication between specialists. Comparing imaging studies without structured reporting caused difficulties in evaluating the progress and severity of the dis-

ease. In 2013, an international working group from the Centres of Excellence in Neurodegeneration published Standards for Reporting Vascular changes on Euroimaging (STRIVE). STRIVE provided a common advisory about terms and definitions for the features visible on MRI as well as with structured reporting of changes related to SVD on neuroimaging [6]. According to STRIVE, the MRI scans, can detect a spectrum of white matter lesions which include: recent small subcortical infarcts, lacunes, white matter hyperintensities, perivascular spaces and microbleeds. Brain atrophy is another pathology observed in the course of SVD [6, 7]. Reviewing diverse imaging findings can sometimes be difficult for human eye to decide where the edges of the disease are and therefore to accurately monitor the disease and assess its severity. To help in this case, automatic solutions come to the rescue, which can provide expert's support and therefore speed up establishing patients diagnosis and providing appropriate treatment.

In this paper we highlight the challenge of creating an automatic tool for diagnosis of SVD and present some preliminary results regarding this topic. Moreover, together with medical practitioners, we have tried to cast this problem into the trustworthy AI framework so that the achieved analysis results can be accepted by the medical practitioners and applied to their practice.

2 Diagnosis of small vessel disease - fundamentals

SVD is a blanket term for lesions which appear in imaging studies secondarily to damage of the small vessels endothelium. Their descriptions along with a graphical representation (Fig. 1) are given below.

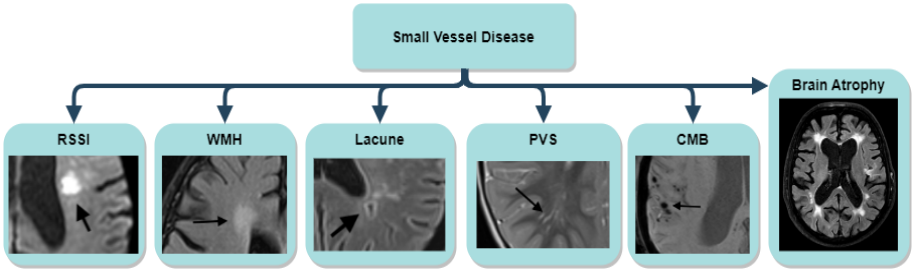

Fig. 1. Samples of SVD lesions

Recent small subcortical infarcts (RSSI) account for almost 25% of all ischemic strokes [5, 8]. RSSIs occur in the areas supplied by a single perforating artery, which are devoid of collateral circulation [6, 9]. According to STRIVE they are best identified on DWI and their diameter is usually smaller than 20mm. On DWI RSSIs are hyperintense focal lesions which strongly restrict water diffusion. They are also hyperintense on T2 and FLAIR. RSSIs are often not visible on CT. Over time, they can evolve into white matter hyperintensities or lacunes, but complete regression is also a possibility [6, 10].

White matter hyperintensities (WMH) are sometimes referred to as leukoaraiosis. They are usually symmetrical, variable in size lesions, which are hiperintense on T2 and FLAIR and isointense or slightly hipointense on T1. Aetiology of WMHs differs depending on location, which can be either periventricular white matter (PVWM) or the deep white matter (DWM). Lesions in PVWM were found to be the result of one of the following: ependymal loss, differing degrees of myelination in adjacent fiber tracts or cerebral ischemia with associated demyelination, whereas DWM lesions are ischemic in nature and their size corresponds with the increasing severity of tissue damage. To quantify the severity of WMHs Fazekas scale is used. It divides WMH lesions based on their location (either PVWML or DWML); lesions located in each of the areas receive a grade from 0 to 3 based on the size of lesions [11].

Lacunes are oval or round lesions which appear at the location of previous small subcortical infarcts or less frequently, microbleeds. They are hypointense on FLAIR and T1 and their diameter is 3-15mm. Usually, lacunes have a hyperintense rim on FLAIR, which allows us to distinguish them from the perivascular spaces [6, 10].

Perivascular spaces (PVS) are also known as Virchow-Robin spaces. They are fluid-filled spaces which surround arterioles, capillaries and venules in the brain. Once enlarged, PVSs become visible in imaging studies as a linear or round hypointense lesions on FLAIR and T1 with basal ganglia being the most common location. Rarely, they can become significantly enlarged and form tumefactive perivascular spaces which can cause mass-effect on surrounding brain tissue [12]. Although, PVSs are normal anatomical structures, it was observed that their number and size increases with the patients age and appearance of other lesions associated with SVD. It was also reported, that there is an association between PVS's and subsequent onset of dementia [13].

Cerebral microbleeds (CMB) are lesions which are hipointense on T2*/ SWI sequences and isointense on T1, T2 and FLAIR. They are usually round or ovoid and smaller than 10mm. The size of microbleeds may vary due to "blooming artifact" which may cause cause micro bleeds to appear larger than they actually are. Differential diagnosis includes intracranial calcifications, metastases susceptible to bleeding and diffuse axonal injury [6, 10].

Brain atrophy is a common outcome of the disease process which affects brain parenchyma and it can be either focal or generalised. Evaluating the change in the size of the brain might be a valuable tool in monitoring the progress of the disease [6, 10]. It is important to note the variety of these lesions in terms of their nature, appearance, number of occurrences, and how they are diagnosed and described by medical specialists. Different sequences such as SWI or FLAIR are used to recognize them and different numbers of examinations are performed to visualize the changes over time as in the case of brain atrophy. In addition, with some lesions such as CMB the valuable information is the number of lesions, while with brain atrophy it is its volume. Due to the factors described above, in order to make a reliable diagnosis, many principles and standards must be considered.

Table 1. Simple and amended SVD score based on [14].

MRI feature	Quantity	Simple SVD score	Amended SVD score
Microbleeds	≥ 1	1	1
Lacunes	0	0	0
	1-2	0	1
	3-5	1	2
	>5	1	3
White matter hyperintensities (Fazekas score)	0	0	0
	1	0	1
	2	1	2
	3	1	3
Total SVD score (range)		0-3	0-7

Two scales, simple and amended SVD score, were developed to diagnose SVD, they are visible in Table 1. In the case of both scales, any score higher than 0 indicates SVD. The higher the score, the more severe SVD is. It is obvious that with such a complex and complicated issue, it is not easy to provide a clear-cut diagnosis. Such diagnosis should be evaluated through the use of various diagnostic tools and consultation with other specialists.

One of these tools may be AI algorithms, however, because of the diversity in how each SVD component is evaluated, it is not possible to use a single universal classifier for this purpose. A decision system that comprehensively diagnoses SVD and assesses its level of severity must consist of an assembly of multiple ML algorithms and a system that draws final conclusions based on it, such as a fuzzy inference system, or a neural classifier.

3 The need for trustworthiness of AI-based systems

ML-based CDSS systems, despite their many advantages, still have many significant flaws and weaknesses. The fully automated and complex data-driven nature of AI models especially deep learning seems to be a double-edged sword. First of all, the model performance strongly depends on provided training data. Unfortunately, the data are usually affected difficult to avoid bias [15]. Examples of such bias include the type and settings of the imaging machine, reasons for the examination, systematic errors by clinicians, and non-statistical appropriateness of the examination group for a given disease due to age, associated diseases, sex, race, etc. of the patients. Additionally, medical image data are usually high-resolution and multidimensional complex structure, while ML algorithms work with down-scaled input, resulting in blurring or even loss of important details [16]. Another data problem is that the publicly available benchmark data that would allow comparisons of the proposed approaches differ significantly from the raw data e.g. from the hospitals. They are often already prepossessed/balanced in some way or there is no information about the survey cohort and imaging

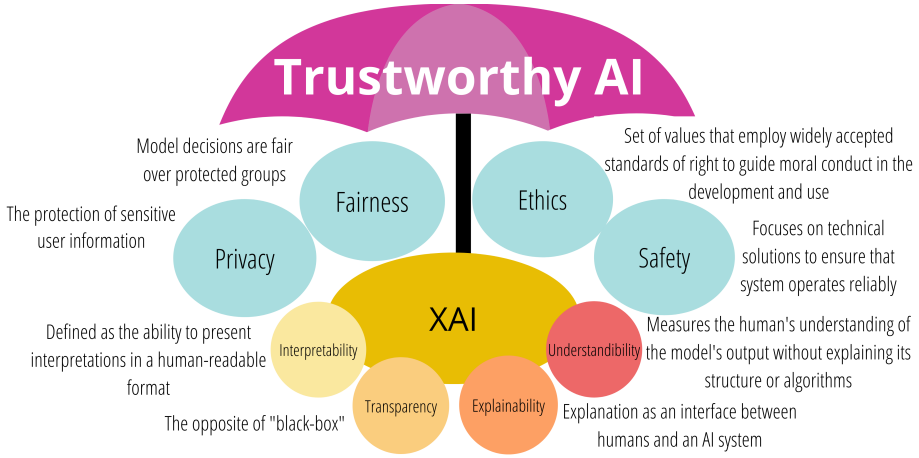

Fig. 2. Terms that are included in trustworthiness

parameters. Furthermore, DNN architectures consist of hundreds of layers and millions of parameters, resulting in a complex black-box model. Additionally, these architectures also, similarly to data, suffer from bias such as evaluation bias, deployment bias or illusion of control bias. Those pose problems related to their robustness, transparency of operation, and ability to generalize, which in turn implies problems related to their trustworthiness. [17, 18].

In the literature, the term trustworthiness includes many different terms. Their definitions are shown in Fig. 2. From the computer science point of view those terms means often something else then for the end users – the clinicians in this case. The developers for example need tools for sanity check the information system and in particularly AI models, mainly in terms of reliability, speed of operation and performance. Medical specialists, on the other hand, need to be shown and explained the links between the features extracted by AI algorithms from the data and their decisions, often giving less attention to the accuracy. It is important to pay attention to reliability of such a system because it might gain trust from the end-users, so it is a key driver for its deployment in clinical practice [18]. When thinking about the CDSS, an important question to consider is what can be done better by AI and what can be done better by the (human) clinician? Simple but time-consuming and tedious work should be left to the algorithms to not waste clinicians time, whereas complex, uncertain problems still require human expertise. In this regard, a reasonable approach is a fusion of both – interactive machine learning with a "human in the loop" that would combine the conceptual understanding and experience that clinicians have and automation of simple tasks. However, the simplicity and intuitiveness of CDSS should be taken into account so that the collaboration between the system and the clinician does not consume the clinician's mental resources [16].

The principle of non-maleficence in the context of medical, which states that clinicians have a fundamental duty not to harm their patients either intentionally

or not, creates the need for explainability. This is necessary in the context of using these black-box models. In such vulnerable and fragile fields as medicine any mistakes, especially false negative results, are a significant problem. The ability of AI to explain its results enables disagreements between the system and human experts to be resolved. It allows the latter to make an informed decision whether or not to rely on the system's recommendations and to provide proper treatment, consequently increasing their trust in the system [17].

What is more, as a result of legal and forensic constraints, such as the European Union General Data Protection Regulations (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the U.S. Food and Drug Administration (FDA), it is not allowed to use AI systems as black boxes [18, 16]. The clinician must be able to understand why a certain decision has been reached. Due this future human AI interfaces should focus on explainability and interpretability.

A relatively new but rapidly growing field of AI addressing these problems is Explainable Artificial Intelligence (XAI). XAI covers a wide range of features, that can be analysed and explained in different ways. One of them is post-hoc explainability techniques, that suites bests for making AI-based medical tools more transparent and understandable for medicans not familiar with software and AI nuances. Post-hoc explainability include textual explanations, visual explanation, local explanations, explanations by example, explanations by simplification, and feature relevance explanation. These techniques can be applied to existing systems without changing their structure, algorithms, or other features. In our opinion, such techniques should be applied on a wide scale, both to systems already being in use and those being developed, especially in the context of medical applications to meet the described expectations of users and to solve the problems highlighted above - user-centric approach.

4 ML-based system development

Nowadays, systems employing machine learning algorithms, and in particular deep learning are black-box models. That means that we provide an input and get the output, but we are not sure, what happens inside - on what ground the decision is made.

Designing an algorithm for medical purposes is not particularly different from ML algorithms in other fields. It consist of standard steps like: problem definition, data preparation, model design, evaluation and trustworthy assessment. However, it requires special precautions, as it strongly affects a human health and life. In order to have a reliable and robust system, we must make every effort to eliminate as many biases as possible at each stage.

Problem definition: This step requires a comprehensive analysis regarding the nature of the problem. Things that have to be considered are the solved task from the technical point of view e.g. classification, detection, segmentation or other; what data must be provided and in what form, what is the expected output etc.

In order to provide a responsible system it is crucial to consult with the specialists in the field. When the specific problem is concerned it is probable that ML specialist do not have enough knowledge to take into account the whole nature of the problem.

For instance, considering small vessel disease, we assume that we want a 0-100% score describing the probability of SVD presence. However, there are several markers that indicate SVD and each of them must be addressed. In case of cerebral microbleeds the required information is their amount in the brain. Therefore, a detection task is sufficient as each of CMBs must be found and counted. When it comes to recent small subcortial infarcts, lacunes or white matter hyperintensities there is a need for segmentation, because these abnormalities might be much bigger than CMB and have irregular shapes. Moreover, the size is crucial regarding the assessment of the patient's condition and it can not be calculated only based on the detection information. In contrast, a completely different approach should be taken to assess the presence of brain atrophy. This symptom cannot be quantified directly, but relatively to previous examinations. It is usually done by comparison of the white and grey matter volume from two scans.

An important issue at this point is also to make a choice between a 3D and 2D space. Although, MRI images are in 3D form, they actually consist of many 2D images merged together. Current 3D ML algorithms have high computation costs, so using a 2D space seems more suitable. However, it is important to provide information from adjacent slices as they carry valuable information, especially for distinguishing specific lesions from its mimics.

Another challenge is visibility of the lesions on different sequences of one scan. Depending on the lesion characteristic and stage, it is visible on specific sequence. For instance, to distinguish WMH from WML, a DWI sequence is crucial - only WMH will be visible. On the other hand, RSSI will be probably visible only for around 10 days at ADC map. Taking it into consideration, it occurs, that proper SVD diagnosis system design is actually a merge of several independent blocks.

Data preparation: Normally, physicians diagnose SVD based on the MRI image analysis, particularly T2-weighted, T1-weighted and gradient echo/ T2*/ susceptibility-weighted sequences. In such case, the same images should be passed as an input to the system. While human can adapt and change the image properties during the analysis, like for example Gamma value. In the automatic process all the data have to be prepared before the process. It forces a careful data preprocessing, including normalization, brightness and contrast adjustment, resize etc. In case of 2D algorithm, this is also a step of introducing the knowledge from adjacent slices.

Moreover, ML algorithms require properly labeled data for pattern recognition. It is often a problem as data labeling is a very laborious and time-consuming process, however, it is a crucial step as it affects the whole training. Unfortunately, it should be performed by radiologists, as ML specialists simply do not have enough knowledge to create trustworthy annotations. Although it may seem

a simple task, the annotating rules should be agreed: whether the lesion is annotated in every image that it is seen or only in one it is best seen; the level of preciseness; the agreement between several raters etc. Any mistakes in labels, new bias is introduced into the system and it leads to some incorrect features generation. Lack of databases is a major obstacle regarding ML system synthesis. Although medical facilities are in possession of huge amount of data, they are not annotated. To the best of our knowledge, there are no publicly available databases of SVD disease. Even for microbleeds - the simplest type of lesion in this disease - there are only few, small ones. Therefore, even after designing a system, it is hard to compare results with the state-of-the-art. Undeniably, medical databases creation is essential for development of automatic diagnosis.

It is also worth mention, that there are some benchmarks regarding image processing, that enable ML algorithms performance check. They mainly serve for algorithms development and comparison. In case of such specific tasks, they do not apply. Not only because of the problem nature, but also shortage in data and its weaker preparation. Therefore, achieving results similar to those achieved on benchmarks is extremely hard.

Model preparation and regularization: At this stage the actual machine-learning model - or models must be prepared. Firstly, the decision regarding the model has to be made. There is a number of already pre-trained models that can be taken advantage of or alternatively a custom model can be designed. The decision is usually made based on the problem specification. Some models deal better with small objects - like Faster RCNN, other do a precise segmentation and some are faster or have lower computational costs - like YOLO or MobileNet. There is a wide range of architectures for detection [19] and segmentation [20]. At the time of decision, all the features of the model have to be taken into account.

Next, the hyperparameters have to be carefully adjusted. Additionally, some regularization techniques should be applied to improve the system performance. There are plenty of solutions that can be added at this point. The base one is data augmentation - although usefull, should be performed carefully, to not produce images that has no connection with real examples, as it may introduce additional bias.

In case of lack of labeled data, self-supervised learning seems to be a promising approach [21]. There is also a branch of providing a domain knowledge into the system [22], which may be extremely useful, especially in medicine. It is obvious, that patient's medical history and his health state are crucial for diagnosis.

Model evaluation: A proper model evaluation is integral with the system design process. There are a lot of metrics describing the system performance. They have to thoughtfully selected not only to show the results considering various aspect, but also to enable comparison between proposed solutions for a given problem. Different metrics are used for classification - accuracy, sensitivity and specificity; detection - sensitivity, precision, F1 score, mAP; segmentation - pixel accuracy, F1score or mAP.

Depending on the problem additional matrics may be informative and should be considered, like in case of microbleeds detection number of false positive predictions per scan and per one CMB is often reported.

Further important issue is the system balance. For instance, sensitivity and precision should be at the similar level. If sensitivity is relatively high and precision low, there is a high generation of false positive predictions. In other words, system is looking for the exact lesion and does not distinguish it from its mimics. On the other hand, when the precision is high and the sensitivity is low at thesame time, probably not many objects of interest are found. Such events may easily lead to radiologist's lost of trust in the system's performance.

It is worth remembering, that not only providing all the metrics is important, proper interpretation is even more valuable as it can point out some features of the system and be useful for improvement.

Regarding the limitations of training datasets, the good practice is to provide a nameplate with the characteristic of the system and data that it is intended for. It is obvious, that system trained on male Europeans in their 60s will not be satisfying enough in case of Asian women in their 40s.

Model trustworthiness: In case of ML-based systems an inherent step in system's creation is providing its trustworthiness. Although it is an area that still requires a lot of research, consideration of aspects mentioned in Section 3 should be taken. Unfortunately, without this step, even accuracy close to 100% does not make a system able to be clinically used.

5 Small Vessels Disease diagnosis - Preliminary results

Our goal is to develop an automatic, trustworthy system for small vessel disease diagnosis. As described in previous sections it is very complex and challenging problem. One of the main elements of the system being developed is a system for cerebral microbleeds detection [23]. Our solution utilizes the Faster RCNN deep neural network, enhanced by the extra post-processing algorithm for false positive and false negative predictions reduction. The algorithm is based on comparison of predictions from adjacent slices and strongly improves system performance. Our solution achieved 92,62% sensitivity, 89,74% precision and 90,84% F1 score. In the Fig. 3a we present an example of microbleeds detection by the system.

Regarding issues related to the reliability and robustness of the system performance, we conducted a comprehensive study of the factors affecting the model development and learning process. The most important aspect is information from adjacent slices inclusion by merging three one-channel images into one three-channels. It enables using a two dimensional solution for three dimensional problem, whereas 3D neural networks are much more computational demanding. We also found out that, in case of CMBs, applying only one label in the slice, where the lesion is the most visible, is more effective than multiple labels in every slice, where the lesion is visible. Further, a larger image size improves the sensitivity of detection, as the lesion occupies a larger absolute area. Unfor-

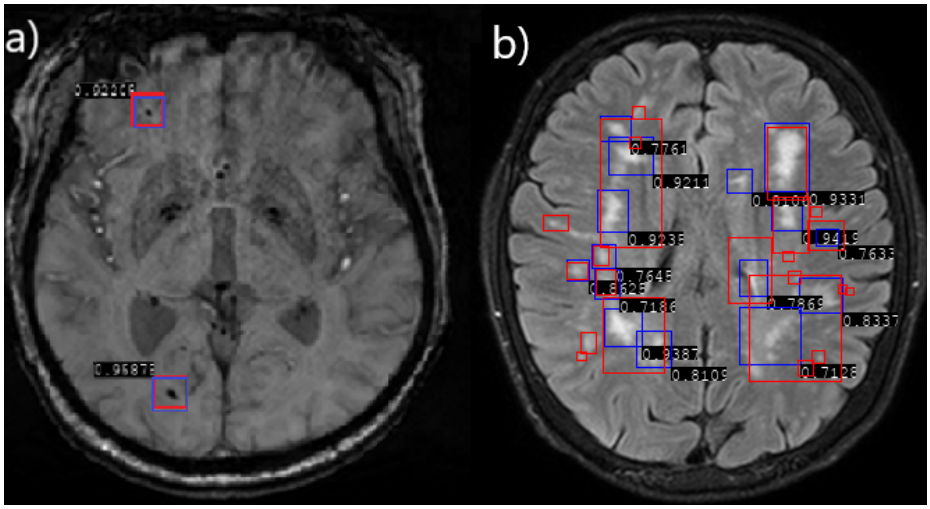

Fig. 3. Samples of detected: a) CMB b) WMH. Ground truth – red, prediction – blue.

tunately, such training has higher computational cost, so the balance between these two aspects must be maintained. Next, by the number of experiments and domain knowledge a proper threshold of prediction confidence score to maintain a balance between sensitivity and precision.

Our current research focuses on a system for diagnosing white matter hyperintensities (WHM). We approached the problem using a similar method to that used for the detection of CMBs. The preliminary results are presented in Fig. 3b. However, it is clearly seen, that detection is not sufficient for this task. Although areas of interest are found with quite good confidence score, any volume and metrics count is inadequate in the current state. One reason is the lack of sufficient amount of labeled data. However, it seems that a much more appropriate approach is to use segmentation instead of detection.

Probably, lacunes, PVS or RSSI also will have to be segmented as WMH, as they have similar geometrical properties, however, different sequences will be considered. While, brain atrophy is a entirely different case - here the volume of brain should be calculated and compared with previous examination.

6 Concluding remarks

The CDSS is first of all designed for clinicians, therefore it should be simple and intuitive in usage, but also reliable and trustworthy.

In recent years, many research report improving metrics of examined algorithms. Obviously, this is an crucial factor regarding AI development. However, discussions with medical specialists suggest that they care more about reliabil-

ity, transparency, and intuitiveness of that systems than about their, sometimes only seemingly, high performance.

Biases are inevitable in ML systems, but all measures should be taken in order to limit their influence. Interpretation of the system and all the rules standing behind enables understanding of decision process making. Moreover, explanation of a specific decision may ease the radiologist work as it not only suggest the diagnosis, but shows the critical areas in the image.

Next essential issue states for results presentation. As long as there are medical regulations regarding disease classification and progression assessment, the system output should be presented in the same way - for example in case of SVD in STRIVE scale. Hence, when designing an applicable diagnosis system we need to pay a special attention to the matter of responsibility if we want it to be used in medical practice.

There are a set of recommendations for ensuring designing of responsible and trustworthy AI systems [24]: use a human-centered design approach; identify multiple metrics to assess training and monitoring; directly examine your raw data, understand the limitations of the dataset and the model; conduct rigorous unit tests to test each component of the system in isolation and as a whole; together with a field specialists design the model using concrete goals for fairness and inclusion; use representative datasets to train and test the model, check the system for biases; analyze performance by using different metrics; treat interpretability as a core part of the user experience, understand the trained model; provide explanations that are understandable and appropriate for the user.

We strongly believe, that matter of trustworthiness still requires a lot of research and methods development, nevertheless it will finally enable wide usage of ML algorithms in medicine.

Acknowledgement

This work was supported by the Ministry of Science and Higher Education in the years 2017–2022, under Diamond Grant DI2016020746.

References

1. J. Bali and O. Bali, "Artificial Intelligence Applications in Medicine: A Rapid Overview of Current Paradigms," *EMJ Innovations*, pp. 73–81, 2020. DOI: 10.33590/emjinnov/19-00167, ISSN: 2513-8634.
2. R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020-02-06. DOI: 10.1038/s41746-020-0221-y, ISSN: 2398-6352.
3. L. Chen, A. L. Carlton Jones, G. Mair, R. Patel, A. Gontsarova, J. Ganesalingam, N. Math, A. Dawson, B. Aweid, D. Cohen, A. Mehta, J. Wardlaw, D. Rueckert, P. Bentley, and For the IST-3 Collaborative Group, "Rapid Automated Quantification of Cerebral Leukoaraiosis on CT Images: A Multicenter Validation Study,"

Radiology, vol. 288, no. 2, pp. 573–581, 2018. DOI: 10.1148/radiol.2018171567, ISSN: 0033-8419, 1527-1315.

4. P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, 2021. DOI: 10.1002/widm.1424, ISSN: 1942-4787, 1942-4795.
5. L. Pantoni, "Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges," *The Lancet Neurology*, vol. 9, no. 7, pp. 689–701, 2010. DOI: 10.1016/S1474-4422(10)70104-6, ISSN: 14744422.
6. J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. DeCarli, F.-E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. v. Oostenbrugge, L. Pantoni, O. Speck, B. C. M. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *The Lancet Neurology*, vol. 12, no. 8, pp. 822–838, 2013. DOI: 10.1016/S1474-4422(13)70124-8, ISSN: 14744422.
7. G. A. Rosenberg, A. Wallin, J. M. Wardlaw, H. S. Markus, J. Montaner, L. Wolfson, C. Iadecola, B. V. Zlokovic, A. Joutel, M. Dichgans, M. Duering, R. Schmidt, A. D. Korczyn, L. T. Grinberg, H. C. Chui, and V. Hachinski, "Consensus statement for diagnosis of subcortical small vessel disease," *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 1, pp. 6–25, 2016. DOI: 10.1038/jcbfm.2015.172, ISSN: 0271-678X, 1559-7016.
8. C. Moran, T. G. Phan, and V. K. Srikanth, "Cerebral Small Vessel Disease: A Review of Clinical, Radiological, and Histopathological Phenotypes," *International Journal of Stroke*, vol. 7, no. 1, pp. 36–46, 2012. DOI: 10.1111/j.1747-4949.2011.00725.x, ISSN: 1747-4930, 1747-4949.
9. J. M. Wardlaw, C. Smith, and M. Dichgans, "Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging," *The Lancet Neurology*, vol. 12, no. 5, pp. 483–497, 2013. DOI: 10.1016/S1474-4422(13)70060-7, ISSN: 14744422.
10. Y. Shi and J. M. Wardlaw, "Upyear on cerebral small vessel disease: a dynamic whole-brain disease," *BMJ*, vol. 1, no. 3, pp. 83–92, 2016. DOI: 10.1136/svn-2016-000035, ISSN: 2059-8688, 2059-8696.
11. F. Fazekas, J. Chawluk, A. Alavi, H. Hurtig, and R. Zimmerman, "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging," *American Journal of Roentgenology*, vol. 149, no. 2, pp. 351–356, 1987. DOI: 10.2214/ajr.149.2.351, ISSN: 0361-803X, 1546-3141.
12. G. M. Potter, F. J. Marlborough, and J. M. Wardlaw, "Wide Variation in Definition, Detection, and Description of Lacunar Lesions on Imaging," *Stroke*, vol. 42, no. 2, pp. 359–366, 2011. DOI: 10.1161/STROKEAHA.110.594754, ISSN: 0039-2499, 1524-4628.
13. J. Ding, S. Sigurðsson, P. V. Jónsson, G. Eiriksdottir, A. Charidimou, O. L. Lopez, M. A. van Buchem, V. Guðnason, and L. J. Launer, "Large Perivascular Spaces Visible on Magnetic Resonance Imaging, Cerebral Small Vessel Disease Progression, and Risk of Dementia: The Age, Gene/Environment Susceptibility–Reykjavik Study," *JAMA Neurology*, vol. 74, no. 9, p. 1105, 2017. DOI: 10.1001/jamaneurol.2017.1397, ISSN: 2168-6149.
14. A. Amin Al Olama, J. M. Wason, A. M. Tuladhar, E. M. van Leijsen, M. Koini, E. Hofer, R. G. Morris, R. Schmidt, F.-E. de Leeuw, and H. S. Markus, "Simple

MRI score aids prediction of dementia in cerebral small vessel disease," *Neurology*, vol. 94, no. 12, pp. e1294–e1302, 2020. DOI: 10.1212/WNL.0000000000009141, ISSN: 0028-3878, 1526-632X.
15. A. Mikołajczyk, M. Grochowski, and A. Kwasigroch, "Towards explainable classifiers using the counterfactual approach – global explanations for discovering bias in data," no. arXiv:2005.02269, 2020. DOI: 10.48550/arXiv.2005.02269.
16. E. Sorantin, M. G. Grasser, A. Hemmelmayr, S. Tschauner, F. Hrzic, V. Weiss, J. Lacekova, and A. Holzinger, "The augmented radiologist: artificial intelligence in the practice of radiology," *Pediatric Radiology*, 2021. DOI: 10.1007/s00247-021-05177-7, ISSN: 0301-0449, 1432-1998.
17. the Precise4Q consortium, J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 2020. DOI: 10.1186/s12911-020-01332-6, ISSN: 1472-6947.
18. Q. Liu and P. Hu, "Extendable and explainable deep learning for pan-cancer radiogenomics research," *Current Opinion in Chemical Biology*, vol. 66, p. 102111, 2022. DOI: 10.1016/j.cbpa.2021.102111, ISSN: 13675931.
19. S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022. DOI: 10.1016/j.dsp.2022.103514, ISSN: 10512004.
20. S. Hao, Y. Zhou, and Y. Guo, "A Brief Survey on Semantic Segmentation with Deep Learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020. DOI: 10.1016/j.neucom.2019.11.118, ISSN: 09252312.
21. A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on Contrastive Self-Supervised Learning," *Technologies*, vol. 9, no. 1, p. 2, 2020. DOI: 10.3390/technologies9010002, ISSN: 2227-7080.
22. D. Schuster, S. J. van Zelst, and W. M. van der Aalst, "Utilizing domain knowledge in data-driven process discovery: A literature review," *Computers in Industry*, vol. 137, p. 103612, 2022. DOI: 10.1016/j.compind.2022.103612, ISSN: 01663615.
23. M. A. Ferlin, M. Grochowski, A. Kwasigroch, A. Mikołajczyk, E. Szurowska, M. Grzywińska, and A. Sabisz, "A comprehensive analysis of deep neural-based cerebral microbleeds detection system," *Electronics*, vol. 10, no. 18, 2021. DOI: 10.3390/electronics10182208, ISSN: 2079-9292.
24. GoogleAI, "Responsible ai practices." https://ai.google/responsibilities/responsible-ai-practices?category=general. Accessed: 2022-05-14.