Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 214 (2022) 18–25

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Information Technology and Quantitative Management

Doctor/Data Scientist/Artificial Intelligence Communication Model. Case Study.

Smaranda Belciug[a]*, Renato Constantin Ivanescu[a], Sebastian-Doru Popa[a], Dominic Gabriel Iliescu[ab]

[a]*University of Craiova, A.I. Cuza Str, no 13, Craiova, 200585, Romania*
[b]*University of Medicine and Pharmacy, Petru Rares Str, no. 2, 200349, Romania*

Abstract

The last two years have taught us that we need to change the way we practice medicine. Due to the COVID-19 pandemic, obstetrics and gynecology setting has changed enormously. Monitoring pregnant women prevents deaths and complications. Doctors and computer data scientists must learn to communicate and work together to improve patients' health. In this paper we present a good practice example of a competitive/collaborative communication model for doctors, computer scientists and artificial intelligence systems, for signaling fetal congenital anomalies in the second trimester morphology scan.

Keywords: deep learning; statistical learning; computer aided medical diagnosis; statistics; second trimester morphology; congenital anomalies.

1. Introduction

How to make a medical decision in the digital era? Can Artificial Intelligence models and Statistics help us better understand the relationship between doctors and computers? The world as we know it rapidly changes. Since the

* Corresponding author. Tel.: +40 729 12 75 74
E-mail address: sbelciug@inf.ucv.ro

1877-0509
10.1016/j.procs.2022.11.143

COVID-19 pandemic surge, the way we practice medicine has changed. The obstetric care landscape has changed. The maternal-fetal care has been left aside, as hospitals were flooded with COVID-19 patients [1-3]. Congenital anomalies (CAs) are the most encountered cause of fetal death, infant morbidity, and mortality [4]. More CAs have been observed in the COVID era, due to the actual infection or due to therapeutic maneuver, [4]. The pandemic made the number of caesarean sections go up, [6]. Diagnosing CAs during morphology scans allows the doctors to discuss in details the infant's prognosis with the parents. The complications that arise from this situation include, and are not limited to procedural risks, morbidity, mortality, and quality of life for all those involved. Unfortunately, there are discrepancies between the pre- and post-natal diagnosis, [7]. These discrepancies are caused by the sonographer's lack of experience, fatigue, fetal movement, time pressure. The mother's features also play an important role in this procedure. For instance, in 50% of the cases, women that have a body mass index over 30 kg/m^2 have their morphology scan misread, [8].

Now, more than ever, we need to use Artificial Intelligence (AI) in computer-medical aided diagnosis. But along with the clear advantages that come with using computers in signaling congenital anomalies or helping physicians establish diagnoses, comes a wave of criticism: how do we explain to the doctor or the patient the diagnosis set by the computer? How do data scientists make people trust AI?

The relationship and communication between data scientists and doctors are tricky. Rumors of doctorless hospitals make them even harder. The goal of this paper is to present a good practice study case which regards a multi-doctor plus AI decision support system in signaling CAs during second trimester morphology ultrasound. Besides the obvious need of working with open-minded doctors, two key features are needed: a competitive/collaborative system for both humans and computers that set the diagnosis, and statistics to determine whether the result is trustworthy or not, [9, 10]. The question: why should we trust the computer's decision? can be answered quite simple, through another question: why should we trust the doctor's decision? The more experienced the doctor is the more we trust its decision. The same principle should be applied to AI also.

The cons that regard AI in the healthcare systems are summarized below, [11]:

- Current algorithms might have been designed and built using flawed and/or incomplete design specifications.
- Are dependent on untrustworthy software/hardware.
- Can have errors and bugs.
- Are dependent on the context.
- Might introduce failure modes by changing the manner in which doctors work.

The success of AI applied in healthcare projects strictly depends on the collaboration between data scientists and doctors. This relationship is hard to build and maintain because the teams that are involved are multidisciplinary. The differences between team members can lead to the derail of the project. It is a common fact that in literature exists plenty of research involving AI in the medicine, yet there is a disparity between the number of published articles and the clinical applications put to practice. A possible cause might be a lack of clinicians that understand AI and are involved from the onset of the projects, [12]. The failure of research projects is directly linked to the miscommunication between the people involved in the project. As proof we have the successful implementation of AI in industry. The challenge is the communication across totally different disciplines, [13]. Do doctors ask the right questions? Do data scientists answer the right way, and vice versa?

Different studies are trying to establish what is the optimum level of mutual understanding between data scientists and doctors, [14, 15, 16]. It is obvious that a data scientist should not expect a medical doctor to have knowledge regarding model development if understandable terms are not used in the communications. The same principle is needed when a doctor tries to explain medical subtleties that might affect the AI system. Mutual understanding is essential. In the research team each member needs to know her/his role and importance in the team. This gives the project a solid start, [17].

To paraphrase John Gray's book: are data scientists from Venus, and doctors from Mars? If so, can we establish a good communication between them so that AI will move faster in healthcare?

In our proposed model, we work very closely with doctors, and we use Statistics to validate both their decisions and the AI models' decision, building thus trustworthy intelligent decision systems.

The remainder of this paper is organized in 4 sections. Section 2 presents the design of the competitive/collaborative communication model for both doctors and AI system. Section 3 presents the risk management of the model. The paper ends with Section 4 that deals with the conclusions.

2. Competitive/collaborative communication model

The interaction between doctors and data scientists is crucial. Data scientists must find means to better explain why an AI system took a certain decision, so that the doctors embrace it without second thoughts. Data scientists do not have the capacity to change the way doctors think, and as some of them find AI a menace to their jobs, we need to establish effective communication models. In our case study we deal with a multi-participant type of decision-making system, which follows these steps, [18]:

1. The *preparation* step which involves defining the problem, its domain, the state-of-the-art, criteria, constraints, and the decision unit.
2. The *collective understanding* step which involves finding a common path to understanding the problem at hand, and on implementing the decision process.
3. The *solution generation* step which involves designing the decision-making process together with alternatives.
4. The *negotiation and confrontation* step in which the participant discuss their viewpoint and try to convince the others to support their ideas. In a way we can see that this step resembles to the competitive phase of our proposed model, the difference being that in our model the negotiation is done after a statistical hierarchy is established and each participant has a voting weight which is proportional to its performance.
5. The *decision* step in which the most voted idea is selected. This is the collaborative phase in our model.
6. The *monitoring* step in which the decision process is observed so that if any problems should appear, they would be caught at an early stage. For this step we have used in our model operational research techniques, to bring the decision process from the lab to the real world.

The architecture of our project communication model is presented in figure 1.

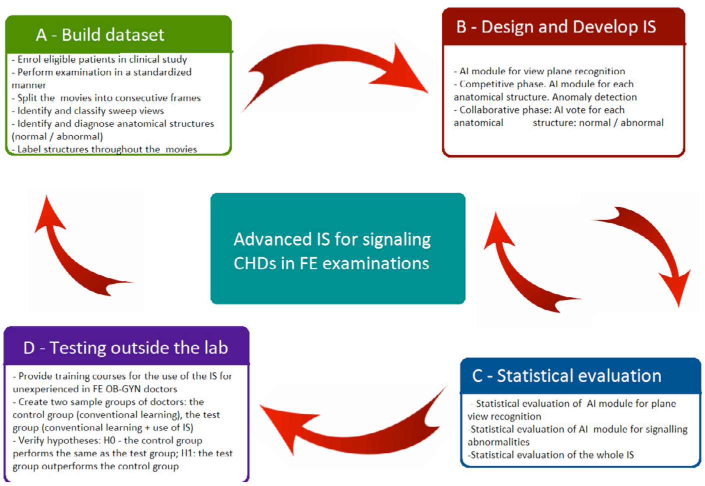

Fig. 1. Architecture of the project communication flow

During the design and implementation of such a model, a set of scientific and technical challenges can be identified. In what follows we shall refer strictly to our study case, using an AI system to signal CAs.

2.1. Cross-disciplinary approach. Understanding the core of diagnosing CA using the morphology second trimester scan requires cross-disciplinary methodologies, that use knowledge from obstetrics and gynecology (OB-GYN), image processing, programming, data base management, statistics, and operational research. This makes the communication between the two teams of doctors and data scientists a truly exciting disciplinary challenge.

Experimental challenges. The hardest experimental challenge that we have encountered in our experiment was the creation of the dataset. We needed to identify and label each visible anatomical structure of the fetus throughout the entire movie. The correct identification of fetus parts is the foundation of which further theoretical and practical developments can be built and validated. For our model we have used the *close* collaboration form. The members of the two decision groups OB-GYN and IT exchange information and make certain decisions [19]. So far, over 400 retrospective ultrasound movies have been reviewed, and over 4000 image frames have been obtained. Patients are enrolled on a daily basis if they agree to be part of the study and they fit the criteria (second trimester of pregnancy). The process of building the dataset started with ultrasound and human anatomy crash course held by the OB-GYN for the computer scientists. In figure 2, which illustrates the learning curve of the computer scientists regarding recognizing fetal anatomical structures on an ultrasound image, we can see that the learning hits a plateau phase, after analyzing over 20 ultrasound movies.

This step was followed by screening each morphology scan and manually classifying the frames into anatomical structures and view planes. After the doctors finished this step, the computer scientists had to anonymize and secure the data, eliminating from the medical images any data that could lead to the identification of the patient. In this step we have used CV2 and Keras-OCR to remove the text from the images. We have used Optical Character Recognition (OCR) to detect the text inside the images, followed by inpainting, to fill the missing parts of the photo in order to produce a complete image.

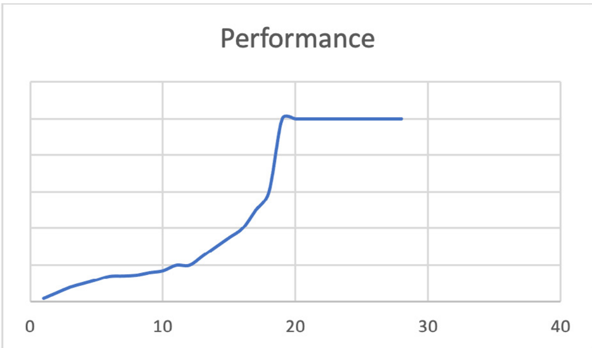

Fig. 2. Learning curve for data scientist regarding fetal anatomical structures on ultrasound movies.

To erase the text from images we had to identify the text and obtain the bounding box coordinates of each text using Kera-OCR. A mask has been applied to each bounding box in order to tell the algorithm which part of the image should be inpainted. In the end the inpating algorithm had been applied to fill the masked area producing a text-free image. The text-free images were then verified by the OB-GYN team to determine if the images were still correct from anatomical point of view, and if we could move further with the building process.

The doctors labeled each structure using LabelStudio. In this stage, at first, they attended multiple LabelStudio crash courses held by some of the computer scientists, and also multiple Python crash courses that enabled them to install Python on their computers, in order to use the annotation program. The OB-GYN doctors consisted of 4 doctors (one senior, and three junior), that worked in a competitive/collaborative manner to label the structures. Each image was labeled by two different doctors, neither of them having access to the other one's labels. After the labeling process was over, the senior doctor established which doctor performed the best at the labeling process and built a hierarchy considering their performances. This is the competitive part of the model.

We have used learning curves to see how the doctors performed after attending the LabelStudio crash course. In figure 3, we can see that after 30 images labeled under the close supervision of the computer scientist, the three junior OB-GYNs started to know how to use this soft.

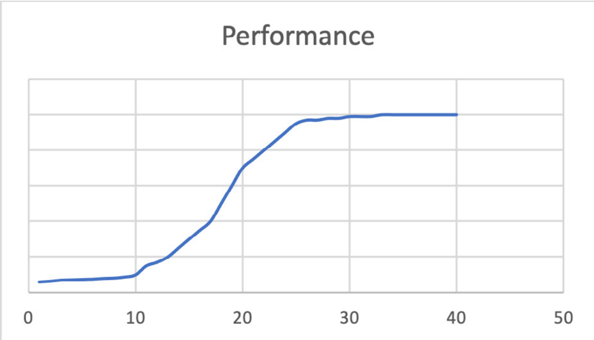

Fig. 3. Learning curve performance after LabelStudio crash course.

We can see here an example of good practice that includes a competitive/collaborative effort doctor/computer scientist in order to achieve a well-built dataset.

Complexity and specificities of the AI System. If the idea of an AI model to perform pattern recognition followed by novelty detection seems rather doable, the number of such structures and the fact that they sometimes overlap is staggering. These features interact between themselves, so, multiple AI models will have to interact also, increasing the complexity of the whole system. Such interactions can take place at several times scales, and both individual interactions and bulk interactions should be taken into account. Each anatomical structure will be detected by an AI model, followed by another model that signals if novelty is found. Just like in the doctors' case, the algorithms that detect a certain structure will be selected through a competition between multiple models, and the 'winner' is chosen statistically.

For this phase, we have used *crowdsourcing*. Technically, we did not assign a certain task to a previously known person/algorithm, rather than anonymous persons/algorithms (crowdworkers) who completed individually the task. The following steps are taken, [20, 21]:

- Defining the corresponding task for solving the decision problem using ideas collected from the crowd.
- Using an open call we broadcast the task to the crowd.
- Let the crowd members propose solutions (classification or novelty task performed by different algorithms).
- Evaluate statistically the results.
- Choose the best solution through a collaborative vote.

2.2. Modelling challenges. All the above-mentioned procedures must be statistically validated to avoid the reproducibility crisis and increase the method robustness.

2.3. From the lab to the real world. The final step of this doctor/computer scientist/AI communication is to test the model in real world situations, which will inevitably face numerous challenges. To verify whether this human-computer interaction is valuable, a statistical comparison is performed between two OB-GYN resident groups that learn to perform the morphology scan in two different manners: (a) *the control group* – in which the residents learn the classical way, observing and learning from an OB-GYN member of the research team, and (b) *the trial group* – in which residents learn the classical way plus using the AI system as a second opinion. The computer scientists will provide training courses to the residents. The evolution of the two sample groups is compared with operational research methods such as learning curves, and different other statistical tools.

The goal of this study case is to:

- *Provide patient access to high quality medical care*. Most women cannot afford antenatal care, and even more they cannot afford morphology scans.
- Have *an experienced sonographer* that knows how to use the AI system, in every small town, even in poor countries, so that we could trust the results of the morphology scans.
- *Avoid misinterpretation* due to time pressure, fetal movement, contractions, fatigue, patient features.

This communication system focuses on three pillars: the committee of doctors and AI models that include deep learning and statistical learning algorithm, statistical analysis, and operational research, [16-18]. The model is high-risk because it includes the communication of people that come from different backgrounds (medicine and computer science) on a very sensitive and emotional subject, but it is also high-gain.

3. Risk management of the competitive/collaborative communication model

To limit the risk and guarantee the success of the doctor/computer scientist/ AI communication, the competitive/collaborative model is designed in a highly iterative approach. This issues in miscommunication that appear during a complete cycle are corrected in the next cycle by modifying each phase accordingly to the lessons learned in the previous step using a Markovian process. The communication was designed to incorporate a gradient risk at several levels, by statistically evaluating the doctors' and AI models' performances. A cycle completes after the statistical comparison of the performances and learning curves of the two sample groups of OB-GYN residents.

The statistical analysis focuses on three aspects: data screening, involving the suitability of the data for the type of statistical analysis that is intended; hypothesis testing, involving the comparison between the testing performances obtained; and over-learning control, including the correlation between the training and testing phases, enabling the statistical investigation of the ability of the model to generalize well new cases. To determine the capability of a statistical test to detect a significant effect, an *a priori* statistical power analysis (two-tailed type of null hypothesis) is performed to determine the appropriate sample size.

The statistical evaluation included normality tests such as Kolmogorov Smirnov & Lilliefors test and Shapiro-Wilk W test. We apply these tests to verify whether the samples are governed by the Gaussian distribution or not. If this is not the case and our sample size is large enough, then we can say that the data distribution is approximately Normal, [19]. The equality of variances was tested using Levene's and Brown-Forshythe tests. We have used these tests, because if the samples have unequal variances, then the Type I error might produce false positives [19, 20]. To differentiate between doctors and AI models we use One-Way ANOVA, and post-hoc Tukey test.

4. Conclusions and future work

The core objective of our paper was to present a good practice example of how doctors and computer scientists communicate in order to make explainable AI. Our model of competitive/collaborative communication system proved that open-minded doctors can be trained to have a little know-how of programming and statistics, whereas computer scientists can be trained in having intel regarding medicine. Even if some doctors are still reluctant to the idea of computer aided medical diagnosis, AI applied in the healthcare system is inevitable, and we should try to find means to better communicate. This system raises the performance of doctors as well as AI models, through competition, and achieves the best results through the collaborative process. None of these results can be possible without the use of the statistical analysis tools.

One thing must be understood and kept in our minds. AI decision systems do not promise a 100% accurate solution to every healthcare problem. They give a solution that worked on a percent of the tested population on that disease, which is exactly what a human doctor does also. In the end, no matter who makes the decision, the doctor, the algorithm, or both, healthcare is a game of luck, in which the patient takes a leap of faith. We need to make this combo work, because there is no competition here, AI works together with the man, not against him.

The slow adoption AI in healthcare may be due to the fact that it has to do with people's health. The fact that AI is successful in other domains such as finance, commerce, entertainment, has to do with the fact that real persons are not involved, and there is no real physical environment. This is why we need to pace things down, and study how AI disruptive application can be integrated in healthcare. After all it is a healthcare revolution. Data scientists need to understand the current restrictions regarding data and accuracy and try to predict the timeline in which things will evolve. Finally, data scientists should not rush the rollout, since we are dealing with people's health.

Acknowledgements

This work was supported by a grant of the Ministry of Research Innovation and Digitization, CNCS-UEFISCDI, project number PN-III-P4-PCE-2021-0057, within PNCDI III.

References

[1] Deprest, J. Choolani, M., Chervenak, F., (2020) et al. Fetal diagnosis and therapy during the COVID-19 pandemic: guidance on behalf of the international fetal medicine and surgery society. *Fetal Diagn. Ther.* **47**: 689-698, https://doi.org/10.1159/000508254.

[2] Mazur-Bialy, A.I., Bogucka, D.K., Tim, S., Oplawski, M. (2020). Pregnancy and Childbirht in the COVID-19 Era – the course of disease and maternal-fetal transmission. *J Clin Med.* **9 (11)b**: 3749, https://doi.org/10.3390/jcm9113749.

[3] Chmielewska, B., Baratt, I., Townsed, R., et al. (2021). Effects of the COVID-19 pandemic on maternal and perinatal outcomes: a systematic review and meta-analysis. *Lancet Global Health* https://doi.org/10.1016/S2214-109X(21)00079-6.

[4] Boyle, B, et al. (2018). Estimating Global Burden of Disease due to congenital anomaly: an analysis of European data, *Archives of Disease in Childhood – Fetal and neonatal edition* **103**: F22-F28.

[5] Khan, M., Nabeka, H., Akbar, S. et al. (2020) Risk of congenital birth defects during COVID-19 pandemic: draw attention to the physicians and policymakers. *J. Global Health* **10 (2)**: 020378, https://doi.org/10.7189/jogh.10.020378.

[6] Dube, R., Kar, S.S. (2020). Covid-19 in pregnancy: the foetal perspective- a systematic review. *Neonatology*. **4 (1)**, https://doi.org/10.1136/bmjpo-2020-000859.

[7] Salomon, L., et al. (2008). A score-based method for quality control of fetal images at routine second trimester ultrasound examination. *Prenat. Diagn*. **28 (9)**: 822-827.

[8] Paladini, D. (2009). Sonography in obese and overweight pregnant women: clinical medicolegal and technical issues, *Ultrasound Obstet Gynecol.* **33 (6)**: 720-729.

[9] Belciug, S., Sandita, A., Costin, H., Bejinariu, S.I. (2021) Competitive/collaborative statistical learning framework for forecasting intraday stock market prices". *Studies in Informatics and Control* **20 (2)**: 135-146.

[10] Gorunescu F., Gorunescu, M., Saftoiu, A., Vilmann, P., Belciug Smaranda. (2011) Competitive/collaborative nerual computing system for medical diagnosis in pancreatic cancer detection. *Exp Sys.* **28 (1)**: 33-48.

[11] Abbott, P.A., Weinger, M.B. (2020) Health information technology: fallacies and sober realities – Redux A Homage to Bentzi Karsh and Robert Wears. *Applied Ergonomics*. **82**: 102973.

[12] Beaulieu-Jones, B., Finlayson, S.G., Chivers, C., Chen, I., McDermott, M., Kandola, J., Dalca, A.V., Beam, A., Fiterau, M., Naumann, T., (2019) Trends and focus of machine learning applications for health research, *JAMA Netw Open,* **2,** e1914051.

[13] Mao, Y., Wang, D., Muller, M., K. Varshney, K.R., Baldini, I., Dugan, C., (2019) Aleksandra Mojsilovic, how data scientists work together with domain experts in scientific collaborations: to find the right answer of to ask the right question? arXiv: 10.1145/3361118.

[14] Meyer, M.A. (2019) Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings, *J Am Med Inf Assoc*, **26**, 383-391.

[15] Famia, R.H. (2022) Clinician data scientists? https://rehoyt.medium.com/clinician-data-scientist-9d82104fdad2.

[16] Baskarada, S, Koronios, A. (2017) Unicorn data scientist: the rarest of breeds. *Programmirovanie,* **51**, 65-74.

[17] Bastian, G., Baker, G.H., Limon, A. (2022). Bridging the divide between data scientists and clinicians. *Intelligence-Based Medicine*, **6**, 100066.

[18] Filip, F.G., (2022) Collaborative Decision-making: concepts and supporting information and communication technology tools and systems. *Int J. Comp Comm Control*. **17 (2)**: 4732.

[19] Ciurea, C., Filip, F.G. (2019) Collaborative platforms for crowdsourcing and consensus-based decisions in multi participant environments. *Informatica Economica* **23 (2)**: 5-10.

[20] Filip, F.G., Zamfirescu, C.B., Ciurea, C. (2017) *Computer Supported Collaborative Decision-Making*. 2017; Springer Cham.

[21] Chiu, C.M., Liang, T.P., Turban, E.(2014). What can crowdsourcing do for decision support? *Decision Support Systems* **65**: 40-49.

[22] Belciug, S. Learning deep neural networks' architectures using differential evolution. Case study: medical imaging processing (2022). *Comp in Biol Med.* **146**: 105623.
[23] Belciug, S., (2021) Parallel versus cascaded logistic regression trained single-hidden feedforward neural networks for medical data. *Exp Sys App.* **170**: 114538.
[24] Belciug, S. (2020) Logistic regression paradigm for training a single-hidden layer feedforward neural network. Application to gene expression datasets for cancer research. *J Biol Inf.* **102**: 103373.
[25] Altman, D.G. (1991), *Practical Statistics for Medical Research*, Chapman and Hall, New York.
[26] Belciug, S. (2020). *Artifitical Intelligence in Cancer: Diagnostic to tailored treatment,* Elsevier.