Evaluating the Effectiveness of Explainable Artificial Intelligence Approaches (Student Abstract)

Jinsun Jung[1,2], **Hyeoneui Kim**[1,2,3*]
[1]College of Nursing, Seoul National University, Seoul, Republic of Korea
[2]Center for Human-Caring Nurse Leaders for the Future by Brain Korea 21 (BK 21) Four Project, College of Nursing,
Seoul National University, Seoul, Republic of Korea
[3]The Research Institute of Nursing Science, Seoul National University, Seoul, Republic of Korea
{jung.nurse, ifilgood}@snu.ac.kr

Abstract

Explainable Artificial Intelligence (XAI), a promising future technology in the field of healthcare, has attracted significant interest. Despite ongoing efforts in the development of XAI approaches, there has been inadequate evaluation of explanation effectiveness and no standardized framework for the evaluation has been established. This study aims to examine the relationship between subjective interpretability and perceived plausibility for various XAI explanations and to determine the factors affecting users' acceptance of the XAI explanation.

Introduction

Extensive study of explainable artificial intelligence (XAI) is now being conducted in the field of healthcare as an effort to overcome the black-box challenges as- sociated with the use of advanced artificial intelligence (AI) techniques such as deep learning and ensemble methods (Payrovnaziri et al. 2020). The right to an explanation for individuals affected by AI-based decisions was recently introduced by the General Data Protection Regulation (GDPR) to address the significance of ensuring the explainability of AI (Bodea et al. 2018). Specific criteria for evaluating the effectiveness of explanations have also been provided by the Defense Advanced Research Projects Agency (DARPA) (Gunning and Aha 2019). However, despite its outstanding performance, no clear standard for assessing and describing AI explainability has been established, thus, real-world use of XAI remains challenging (Payrovnaziri et al. 2020).

Previous studies have examined user satisfaction, trustworthiness, and task performance; however, few studies have examined the factors affecting the user's acceptance of explanation approaches, which is an essential reflection of the user's understanding of AI (Aechtner el al.2022; Merry, Riddle, and Warren 2021). To fill this gap, this study aimed to determine how XAI is interpreted by users of AI-generated results, using the results derived from various XAI techniques. This study also aimed to identify the relationship between interpretability and plausibility of different XAI explanations.

Methods

The Korean National Health and Nutrition Examination Survey (KNHANES) was used in the development of the obesity prediction model. In total, 24,798 individuals over 19 years old were selected and 94 variables were included in the model. The binary outcome variable was created for Body Mass Index (BMI) threshold of 25 kg/m2 and was predicted using XGBoost with 93 predictors, including demographics, a health interview, a health examination, and a nutrition survey. Four explanatory AI techniques — XGBoost feature importance, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanation (LIME), and permutation importance — were applied to explain the prediction results. Evaluation of the results was performed by a focus group using a semi-structured interview and an open discussion.

Seven graduate students working on health science research were invited to join the focus group. An outline of the design of the XAI model for the prediction of obesity, with an introduction of the four types of explanatory mechanisms listed above was first presented by the researcher (JJ). The focus group then identified the XAI approaches that best represented their thinking process with regard to obesity prediction according to the following steps: (1) manually identifying and prioritizing the top ten possible risk factors from 93 predictors; (2) reviewing and ranking the subjective interpretability of XAI approaches with rationales;

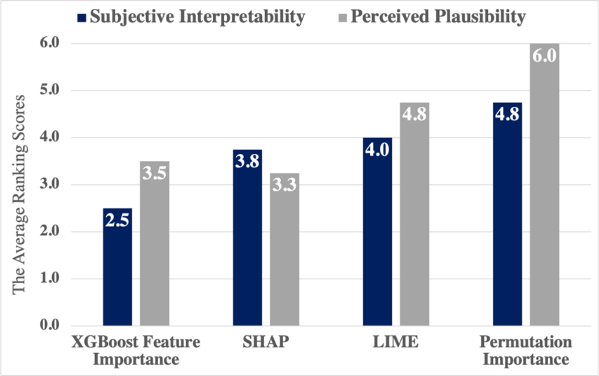

Figure 1. The average ranking scores of subjective interpretability and perceived plausibility

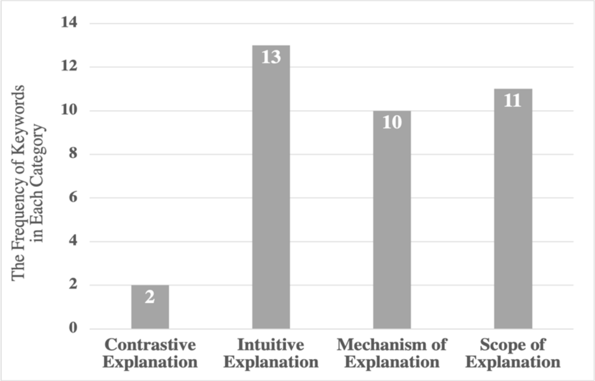

Figure 2. The categorized rationales of the rankings

(3) ranking the XAI generated results according to plausibility; and (4) sharing opinions on explanations of XAI. The rationales presented in step 2 were transcribed during the focus group interview and manually annotated using common keywords and paraphrases, and then categorized into four groups.

Results

Figure 1 shows the average ranking scores for interpretability and plausibility of the four XAI explanation techniques that the focus group provided. Note that a smaller average indicates a high ranking by the focus group. XGBoost feature importance was considered the most plausible XAI explanation, followed by SHAP and LIME. Most of the rankings for interpretability and plausibility remained consistent, except for SHAP.

The justifications of the focus group provided during step 2 for the interpretability ranking of XAI were categorized according to four groups. The intuitiveness of the explanation was most common, followed by the scope of explanation, mechanism of explanation, and contrastive explanation (see Figure 2). The comments recorded during the open-ended discussion suggested that a “why not” explanation (i.e., contrastive explanation) would be helpful in the effort to understand the results and that too many predictors could be a distraction to users (i.e., scope of explanation).

Discussion and Conclusion

A focus group evaluation was used in this study to ex- amine the interpretability and plausibility of the four different XAI approaches. The findings of this study demonstrated that the intuitiveness of the explanation has the greatest impact on interpretability, and interpretability tends to have an association with plausibility. Acceptance of the XAI results was also affected by the scope and mechanism of explanations (e.g., applying sensitive analysis, assigning weights etc.). Presenting contrastive cases can be helpful, although mentioned by only two focus group participants.

Despite the generalizability is limited due to a small scope, the findings of this study suggest that users may consider explanations that are easily interpreted more plausible, which may lead to better acceptance of the associated results. Conduct of a large-scale study including a larger focus group will be needed to establish a standardized framework for evaluating the effectiveness of XAI explanations.

References

Aechtner, J.; Cabrera, L.; Katwal, D.; Onghena, P.; Valenzuela, D. P.; & Wilbik, A. 2022. Comparing User Perception of Explanations Developed with XAI Meth- ods. 2022 IEEE International Conference on Fuzzy Sys- tems (FUZZ-IEEE). IEEE.

Bodea, G.; Karanikolova, K.; Mulligan, D. K.; & Makagon, J. 2018. Automated decision-making on the basis of personal data that has been transferred from the EU to companies certified under the EU-US Privacy Shield: Fact-finding and assessment of safeguards provided by US law. Final report TNO Directorate-General for Jus- tice and Consumers, Directorate C: Fundamental Rights and Rule of Law, Unit C.

Gunning, D.; and Aha, D. 2019. DARPA’s explainable artificial intelligence (XAI) program. AI magazine, 40(2), 44-58.

Merry, M.; Riddle, P.; and Warren, J. 2021. A mental models approach for defining explainable artificial intelligence. BMC Medical Informatics and Decision Making, 21(1), 1-12.

Payrovnaziri, S. N.; Chen, Z., Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J. H.; ... and He, Z. 2020. Explainable artificial intelligence models using real- world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association, 27(7), 1173-1185.