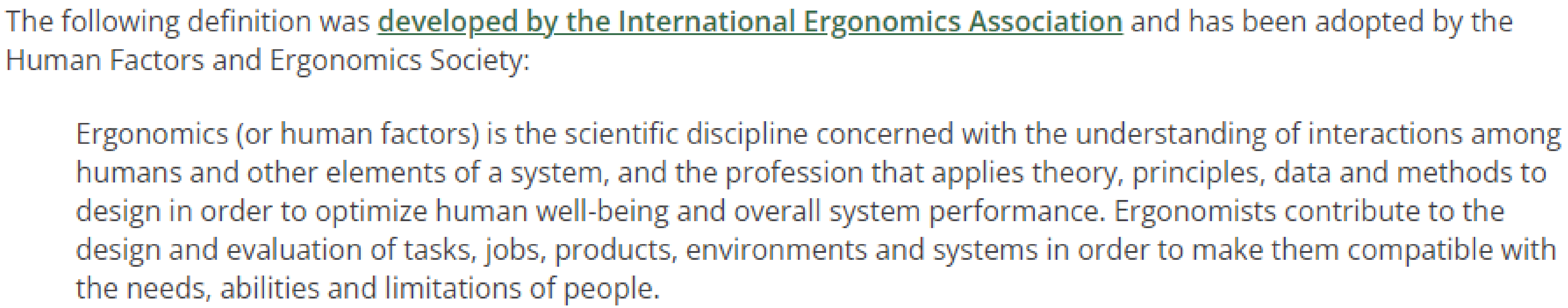
Human factors engineering for explainable AI

Woojin Park, PhD
Life Enhancing Technology Laboratory
Department of Industrial Engineering
Seoul National University

https://www.let.snu.ac.kr/

Human factors/ergonomics



The following definition was developed by the International Ergonomics Association and has been adopted by the Human Factors and Ergonomics Society:

> Ergonomics (or human factors) is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance. Ergonomists contribute to the design and evaluation of tasks, jobs, products, environments and systems in order to make them compatible with the needs, abilities and limitations of people.

Our research topics (last 2 years)

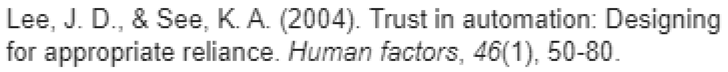
- Human-vehicle system
  - Take-over interface design
- Heavy machinery design
  - Excavator controller design
- Mobile devices
  - Multi-device experience
- Blockchain applications
  - NFT marketplace user experience
- Augmented reality systems
  - AR interface design for supporting manual work tasks
- Foldable displays
  - Perception of foldable display quality
- XAI
  - Smart-chair based low back pain recognition system

Agenda

- To enhance the audience’s understanding of the human factors concept ‘trust’ in the context of human-AI collaboration
- To present our group’s recent work on the comparative evaluation of different explanation types for a smart chair-based low back pain telediagnosis system

Trust in human-AI collaboration

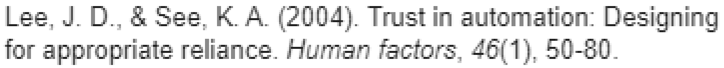
Trust



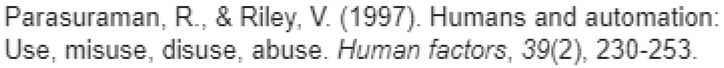
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

- Trust, a social psychological concept, is important for understanding human-automation (AI) partnerships.
- Trust can be defined as: “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.”

Appropriate trust in automation



Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.



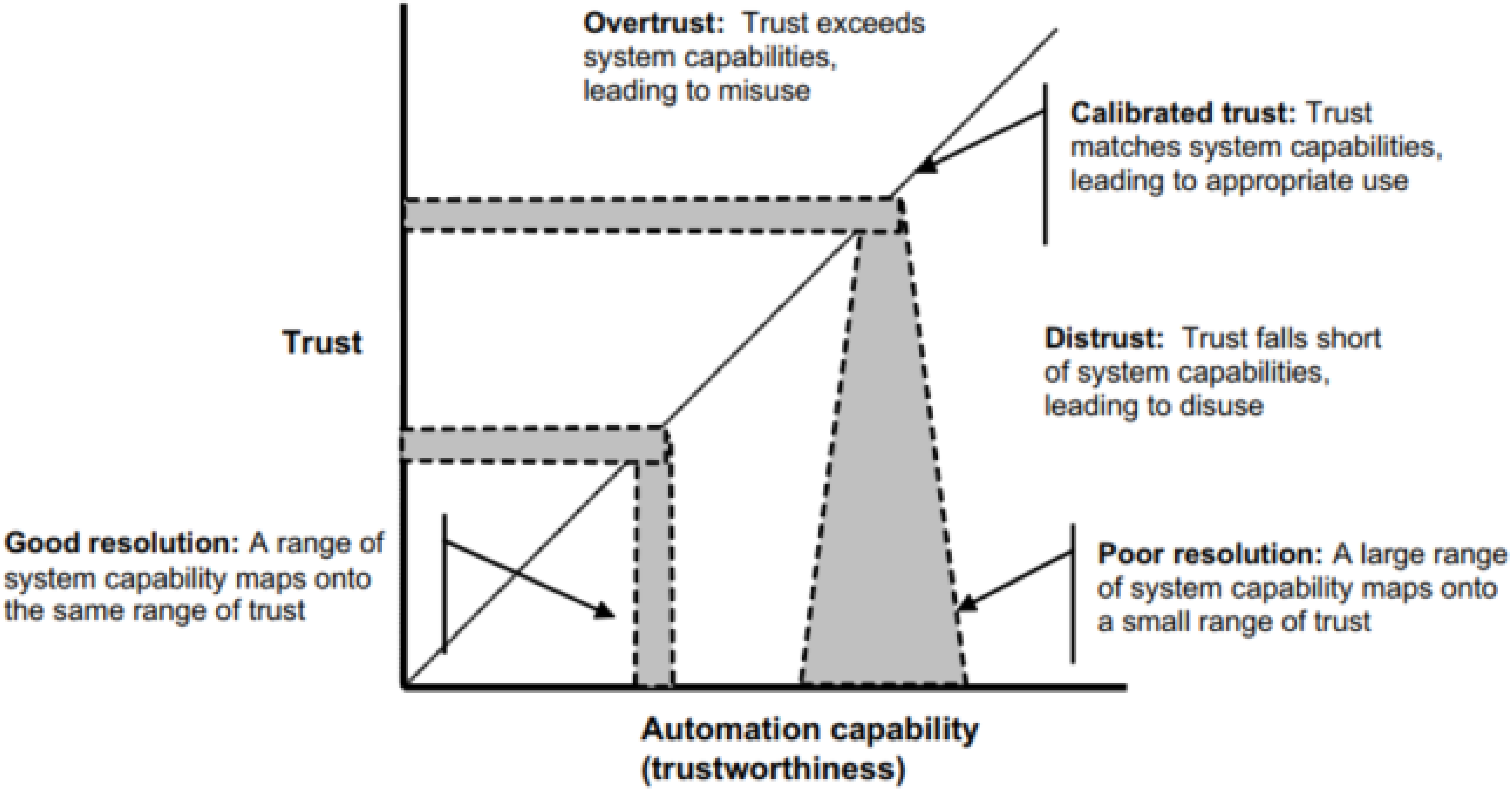
Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

- Appropriateness of trust?
    - ‘The relationship between the true capabilities of the agent and the level of trust’
- Inappropriate reliance associated with misuse and disuse depends, in part, on how well trust matches the true capabilities of the automation.
- Supporting appropriate trust is critical in avoiding misuse and disuse of automation.

Appropriate trust in automation

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.



Overtrust: Trust exceeds system capabilities, leading to misuse

Calibrated trust: Trust matches system capabilities, leading to appropriate use

Distrust: Trust falls short of system capabilities, leading to disuse

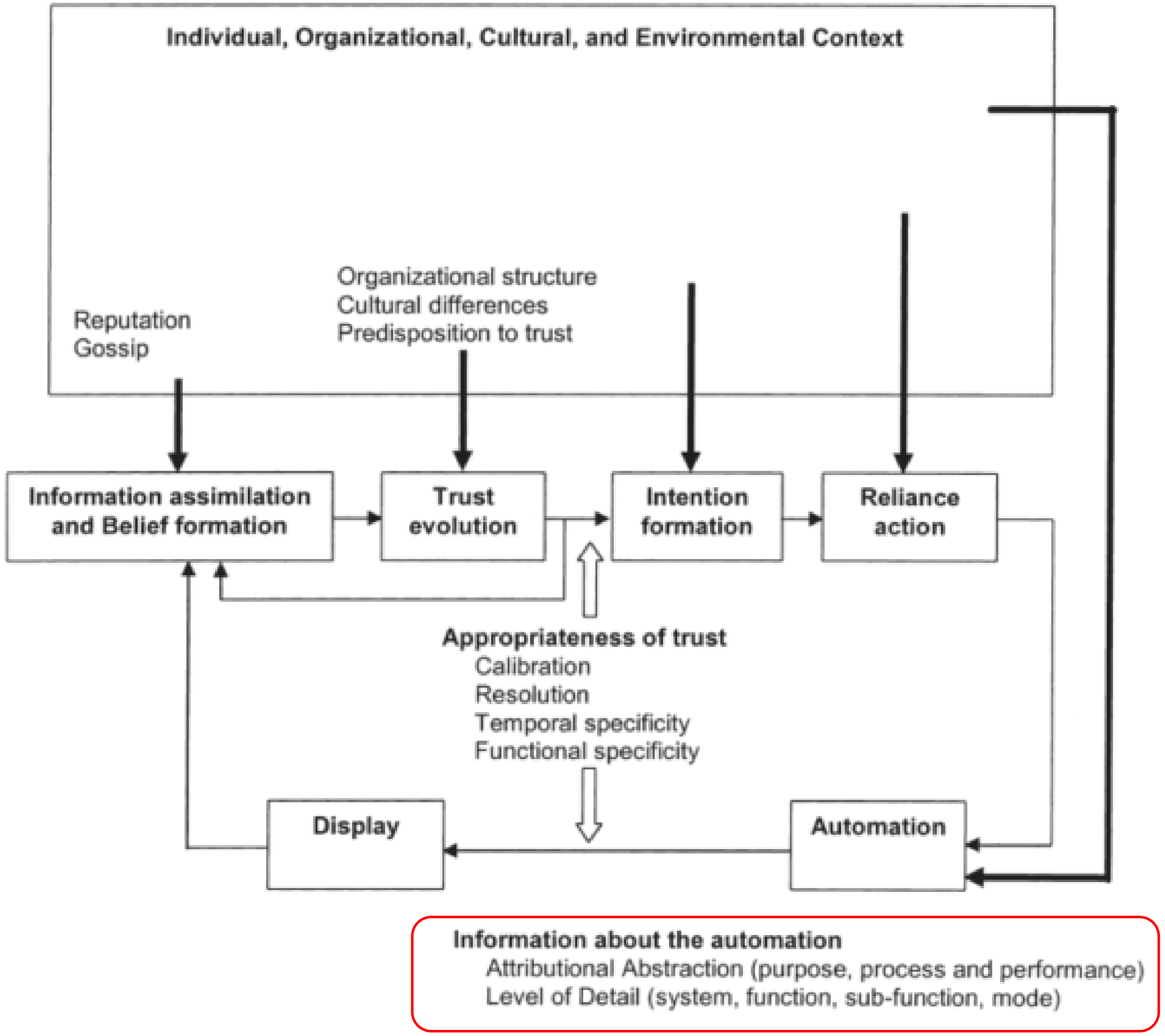
Good resolution: A range of system capability maps onto the same range of trust

Poor resolution: A large range of system capability maps onto a small range of trust

Figure 2. The relationship among calibration, resolution, and automation capability in defining appropriate trust in automation. Overtrust may lead to misuse and distrust may lead to disuse.

Trust

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.



Individual, Organizational, Cultural, and Environmental Context

Reputation
Gossip

Organizational structure
Cultural differences
Predisposition to trust

Information assimilation and Belief formation

Trust evolution

Intention formation

Reliance action

Appropriateness of trust
Calibration
Resolution
Temporal specificity
Functional specificity

Display

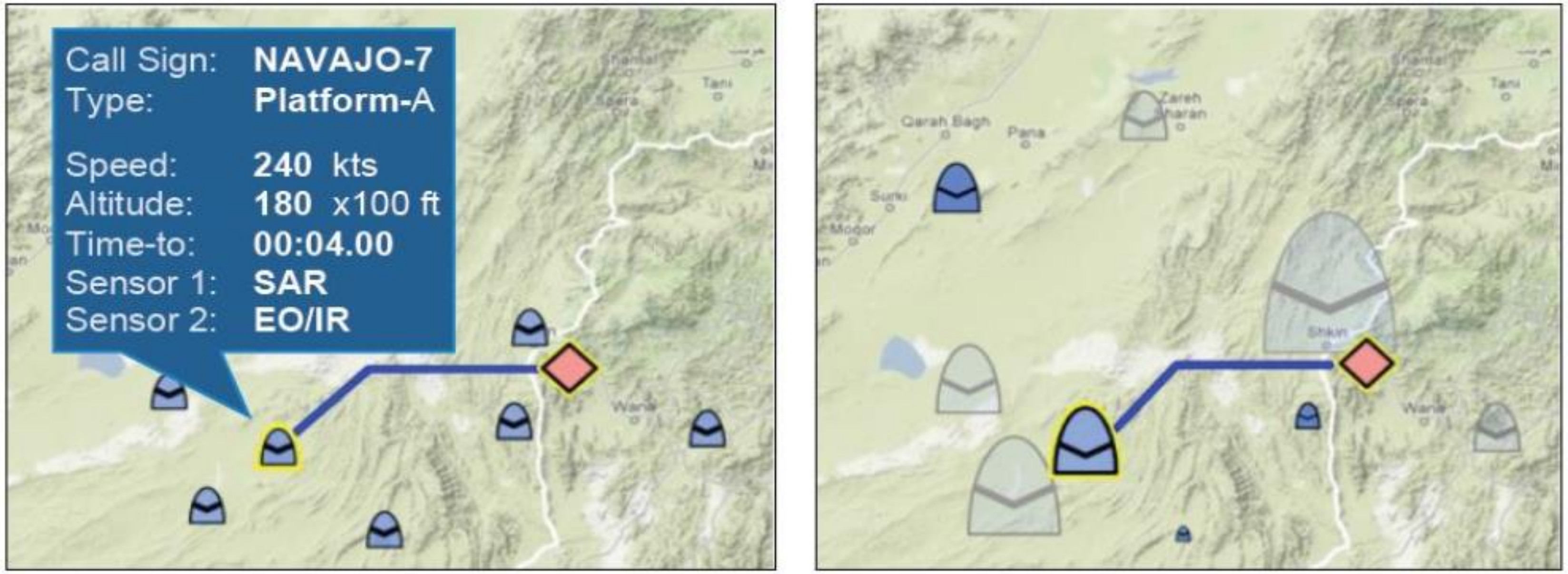
Automation

Information about the automation
Attributional Abstraction (purpose, process and performance)
Level of Detail (system, function, sub-function, mode)

Figure 1. The interaction of context, agent characteristics, and cognitive properties with the appropriateness of trust.

Trust and explainability

- Explanation interface design is a key consideration for supporting appropriate trust.
- Information displays must be designed to explain machine decisions/predictions clearly and in an easy-to-understand manner.



Call Sign: NAVAJO-7
Type: Platform-A
Speed: 240 kts
Altitude: 180 x100 ft
Time-to: 00:04.00
Sensor 1: SAR
Sensor 2: EO/IR

Explanation interface examples

Kilgore, R., & Voshell, M. (2014). Increasing the transparency of unmanned systems: Applications of ecological interface design. In *Virtual, Augmented and Mixed Reality. Applications of Virtual and Augmented Reality: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 6* (pp. 378-389). Springer International Publishing.

Comparative evaluation of explanation interfaces for a smart-chair based low back pain recognition system

Backgrounds

- Domain-specific XAI

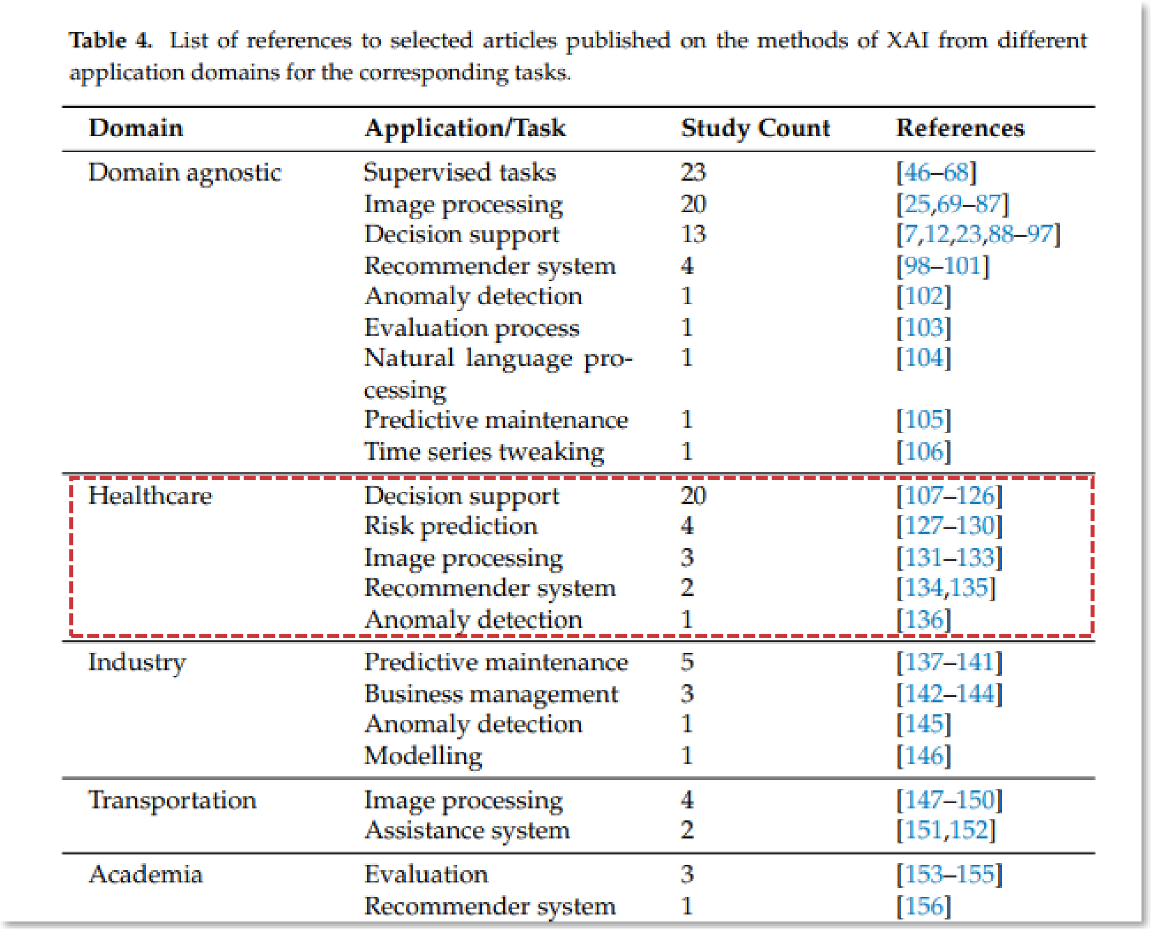


Table 4. List of references to selected articles published on the methods of XAI from different application domains for the corresponding tasks.

Domain	Application/Task	Study Count	References
Domain agnostic	Supervised tasks	23	[46–68]
	Image processing	20	[25,69–87]
	Decision support	13	[7,12,23,88–97]
	Recommender system	4	[98–101]
	Anomaly detection	1	[102]
	Evaluation process	1	[103]
	Natural language processing	1	[104]
	Predictive maintenance	1	[105]
	Time series tweaking	1	[106]
Healthcare	Decision support	20	[107–126]
	Risk prediction	4	[127–130]
	Image processing	3	[131–133]
	Recommender system	2	[134,135]
	Anomaly detection	1	[136]
Industry	Predictive maintenance	5	[137–141]
	Business management	3	[142–144]
	Anomaly detection	1	[145]
	Modelling	1	[146]
Transportation	Image processing	4	[147–150]
	Assistance system	2	[151,152]
Academia	Evaluation	3	[153–155]
	Recommender system	1	[156]

Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the need for explainable artificial intelligence (xAI). *arXiv preprint arXiv:2012.01007*.

Jeong, H., & Park, W. (2020). Developing and evaluating a mixed sensor smart chair system for real-time posture classification: Combining pressure and distance sensors. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1805-1813.

Backgrounds

- Mixed sensor smart chair system for real-time posture classification
  - Classifies a posture as one of 11 predefined sitting posture categories

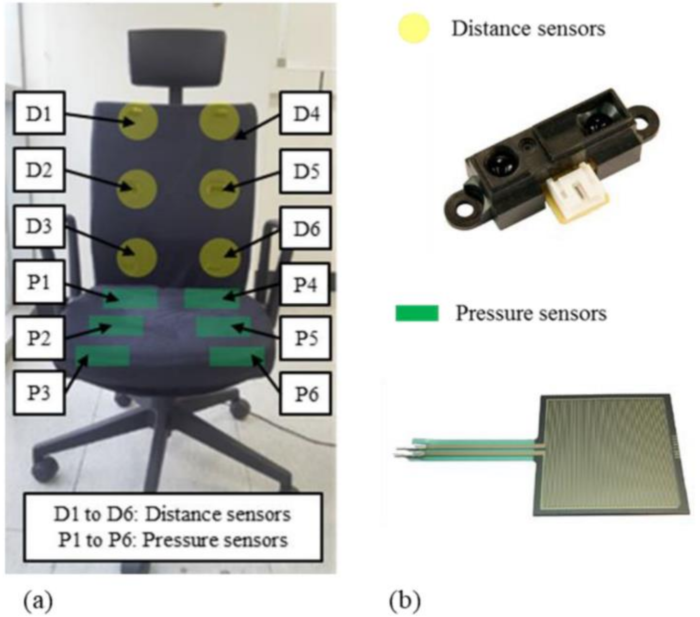
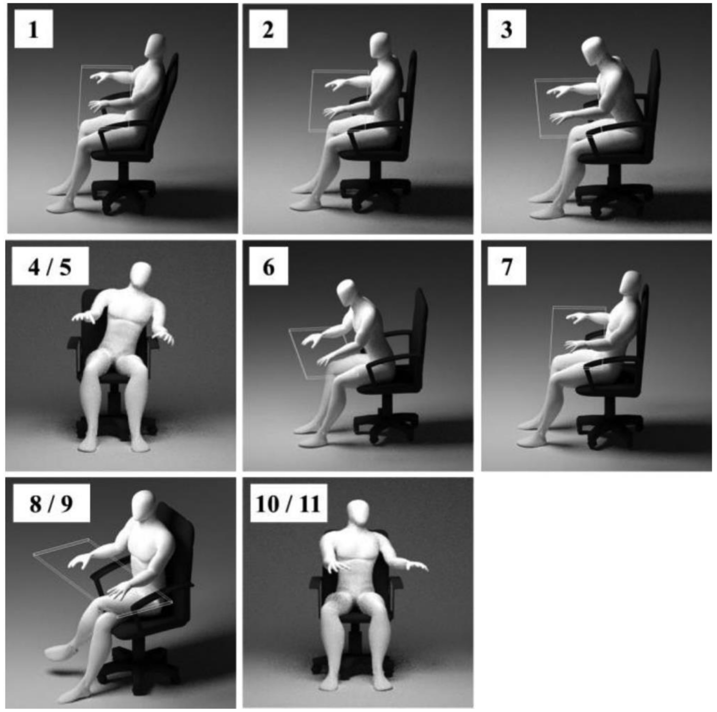


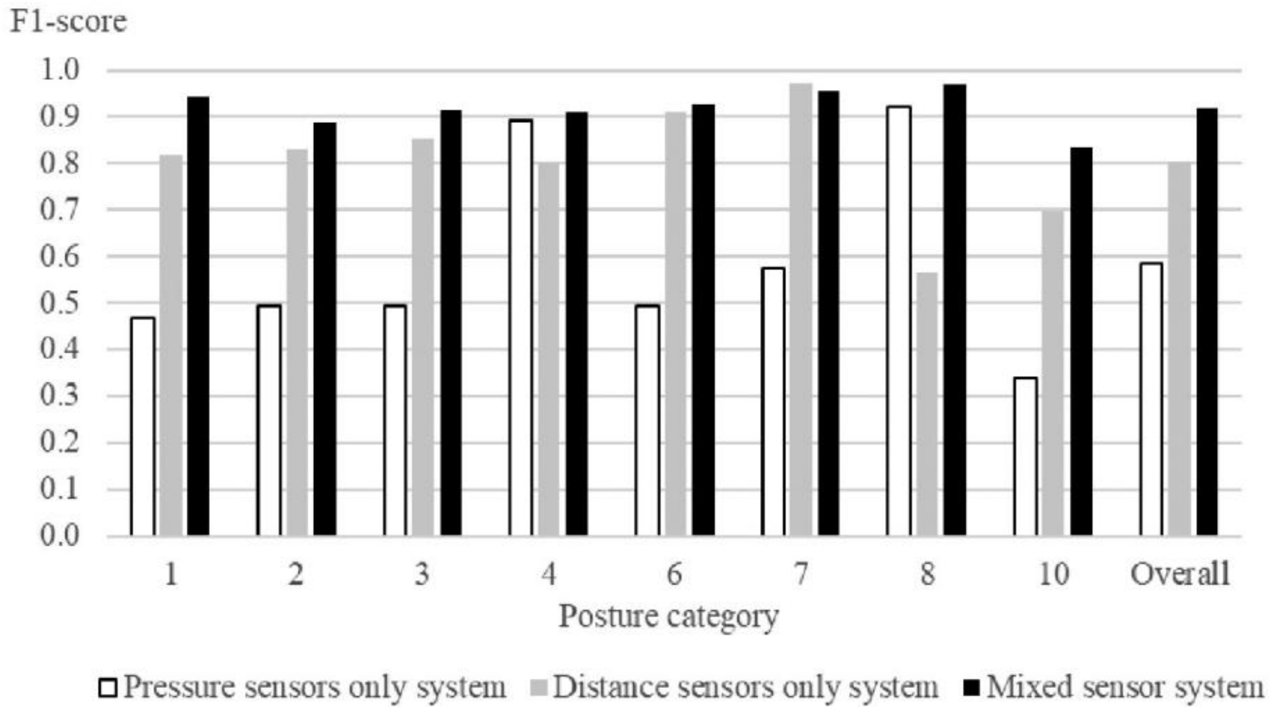
Fig. 2. Physical construction of the mixed sensor system: (a) placement of sensors, and (b) distance and pressure sensors.



Eleven sitting posture categories

Backgrounds

- Smart chair for real-time posture classification



Backgrounds

- Smart chair-based low back pain (LBP) recognition system
  - An extension of the mixed sensor smart chair for real-time posture classification
  - Classifies a user as either a chronic low back pain (CLBP) patient or a non-patient
- LBP
  - The most common disorder worldwide (80% of the population)
  - Causes enormous economic and social losses
  - CLBP refers to LBP persisting for more than 3 months

Backgrounds

- Smart chair-based LBP recognition system
  - The mixed chair smart chair system is used to generate a time sequence of sitting postures while a user is performing computer typing for a one-hour time period.
  - A binary classification model (the CLBP detector) classifies the user as either a CLBP patient or a non-patient on the basis of the posture-time sequence.
    - Machine learning (the CatBoost algorithm) was used to develop the binary classification model.
  - The binary classifier utilizes a set of features:
    - Relative frequencies of some posture categories
    - Time changes in relative frequencies of some posture categories
    - Number of posture categories observed during the one-hour time period

Backgrounds

- Smart chair-based low back pain (LBP) recognition system

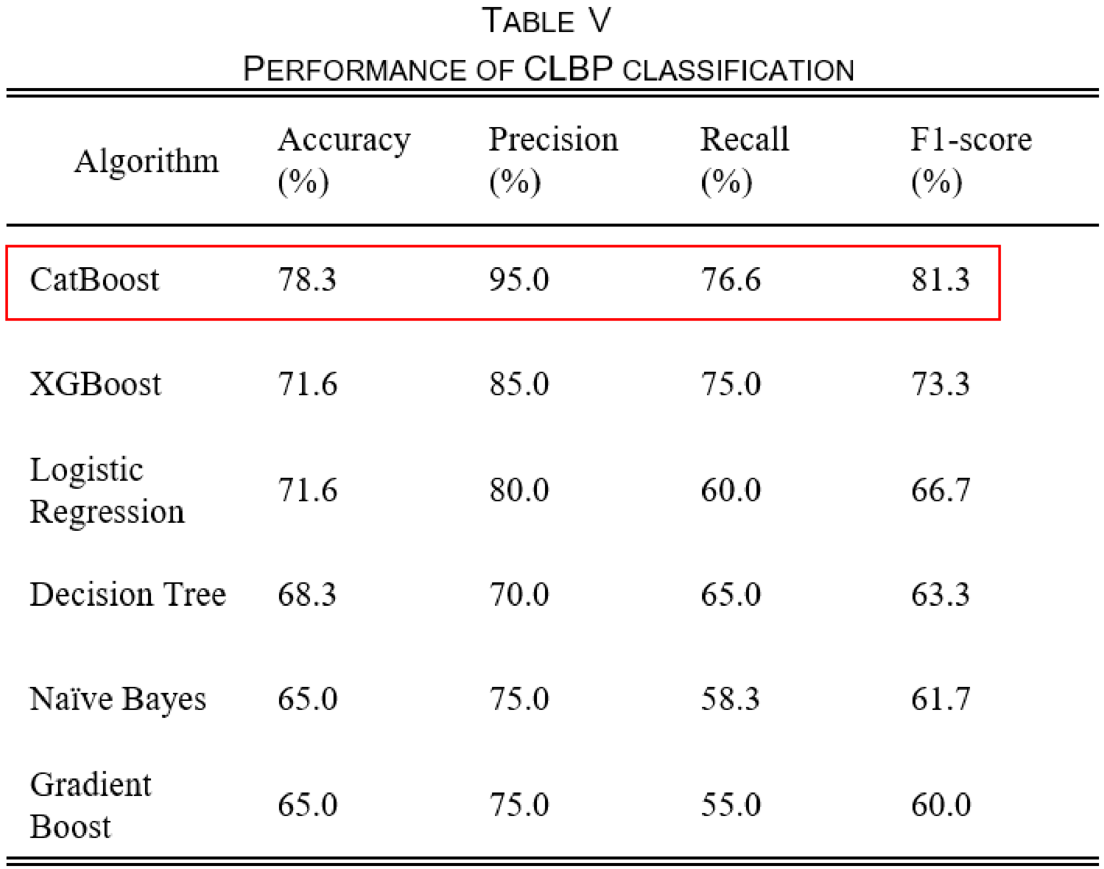


TABLE V
PERFORMANCE OF CLBP CLASSIFICATION

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CatBoost	78.3	95.0	76.6	81.3
XGBoost	71.6	85.0	75.0	73.3
Logistic Regression	71.6	80.0	60.0	66.7
Decision Tree	68.3	70.0	65.0	63.3
Naïve Bayes	65.0	75.0	58.3	61.7
Gradient Boost	65.0	75.0	55.0	60.0

Research problem

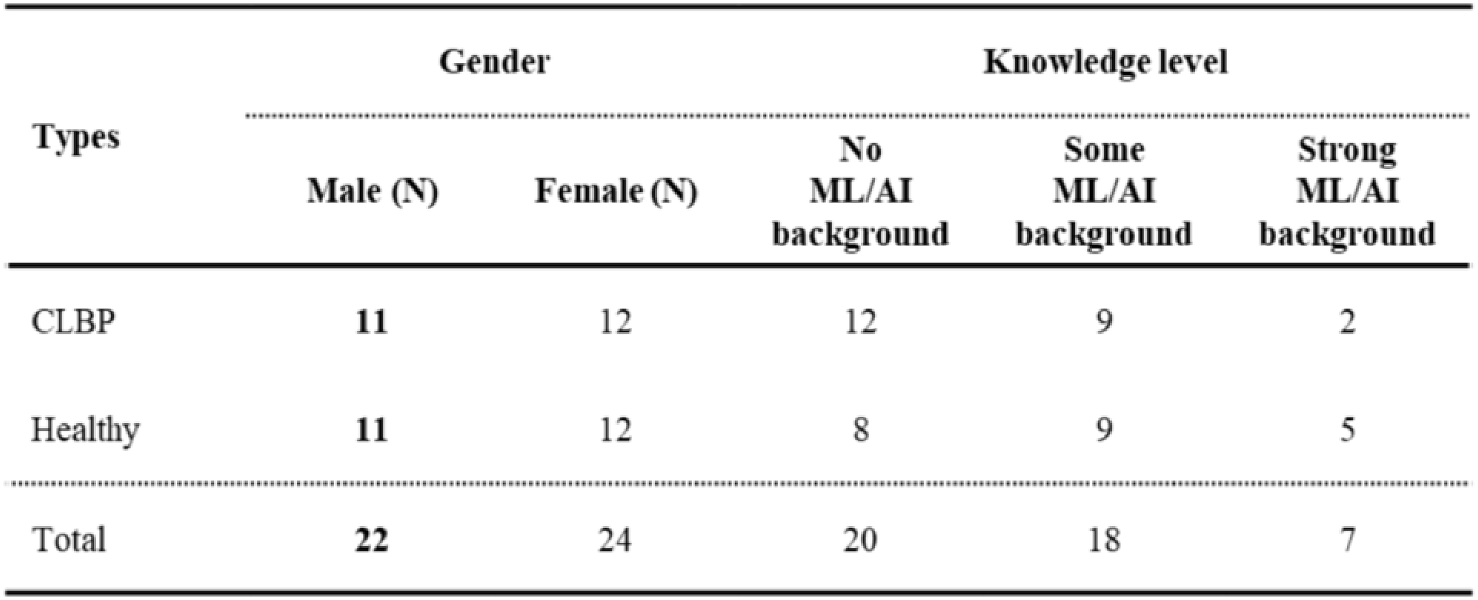
- Research questions:

  RQ 1) How can the CLBP recognition system’s diagnosis result be best explained to the user?

  RQ 2) How do existing XAI methods (explanation types) compare in terms of XAI evaluation metrics?

Experiment

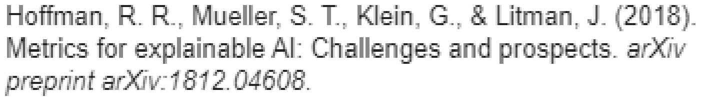
- Participants:
  - 22 males and 24 females
  - 19 ~ 33 years old (24.7 ± 3.4)
  - Each participant was assigned randomly to one of two groups (diagnostic output): the CLBP patient group and the non-patient group
  - Different levels of AI/ML knowledge were represented



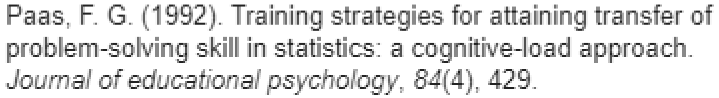
Types	Gender		Knowledge level		
	Male (N)	Female (N)	No ML/AI background	Some ML/AI background	Strong ML/AI background
CLBP	**11**	12	12	9	2
Healthy	**11**	12	8	9	5
Total	**22**	24	20	18	7

Experiment

- Four local explanation types:
  - No explanation
  - Feature attribution explanation
  - Example-based explanation
  - Decision tree explanation
- XAI evaluation metrics:
  - User experience metrics (5-point scales)
    - Understandability
    - Satisfaction
    - Sufficiency
    - Completeness
    - Usefulness
    - Perceived accuracy
    - Trustworthiness
  - Cognitive load (9-point scale)
    - Pass's cognitive load rating



Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.



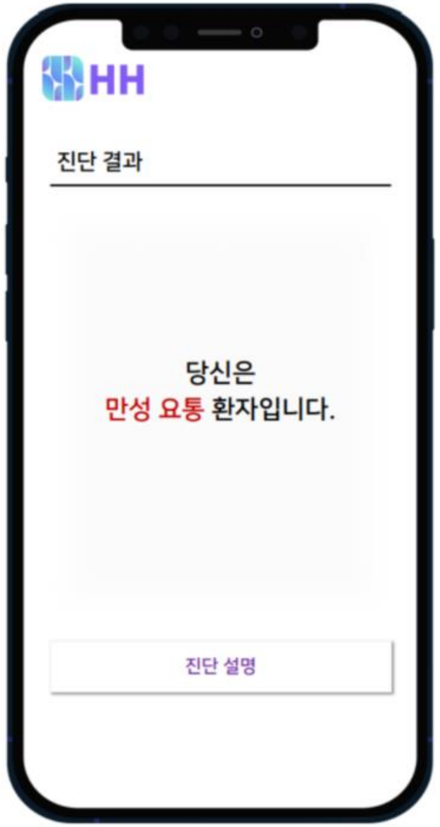
Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology*, *84*(4), 429.

Experiment

- Experimental procedure:
  - Introduction session:
    - The purpose and procedure of the experimental study were described
  - Learning session:
    - The definition and characteristics of CLBP were provided
    - How the smart chair-based CLBP recognition system works was explained
    - The three explanation methods were explained
  - Experimental trials:
    - The interface prototypes for the different explanation methods were presented
    - The participants took enough time to examine and interpret the interface prototypes, and, then, performed subjective ratings (UX and cognitive load)
    - The presentation order of the four explanation types was randomized for each participant, with complete counter balancing

Explanation interface prototypes (diagnostic output: ‘CLBP patient’)

- No explanation

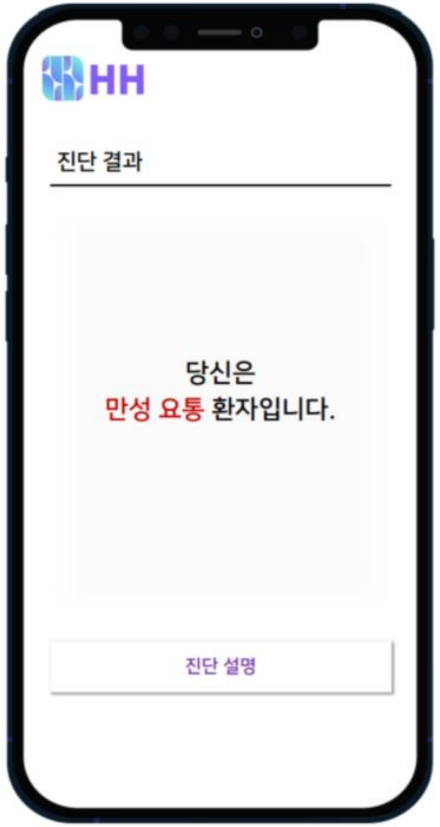


진단 결과

당신은
만성 요통 환자입니다.

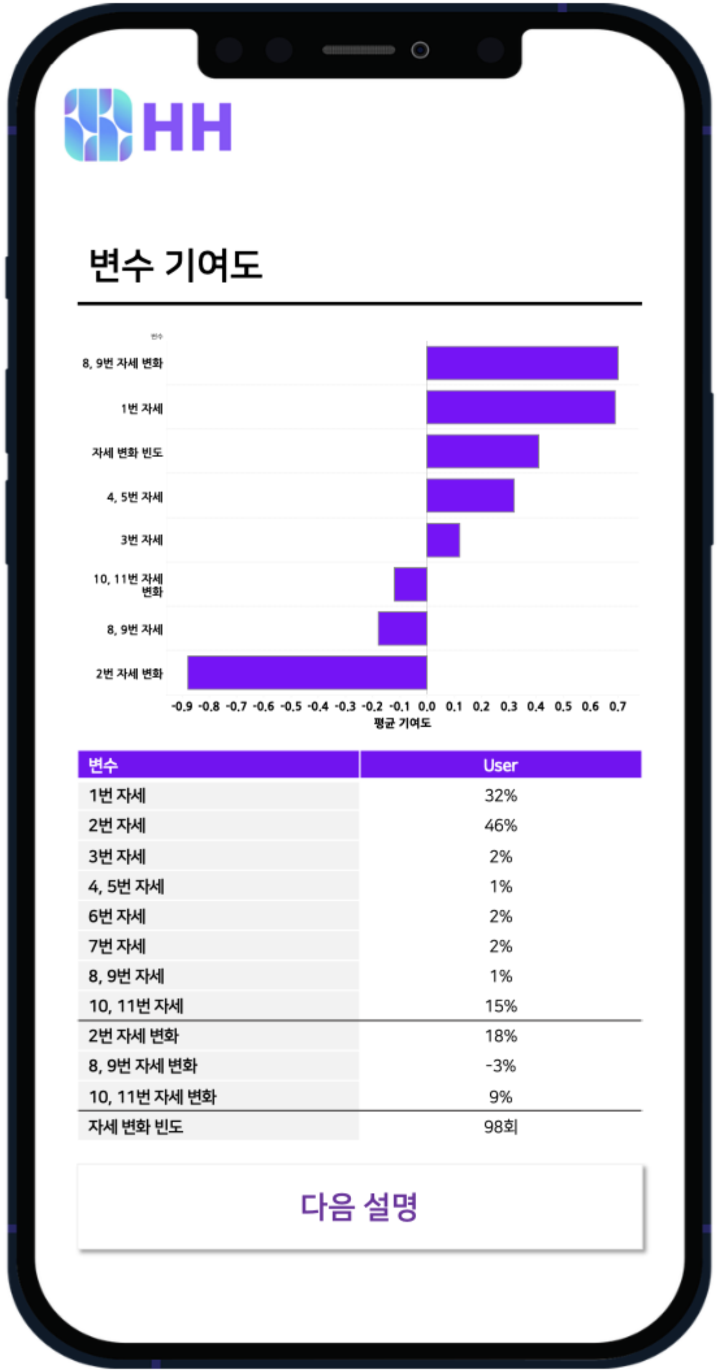
Explanation interface prototypes (diagnostic output: 'CLBP patient')

- Feature attribution explanation



진단 결과

당신은
만성 요통 환자입니다.

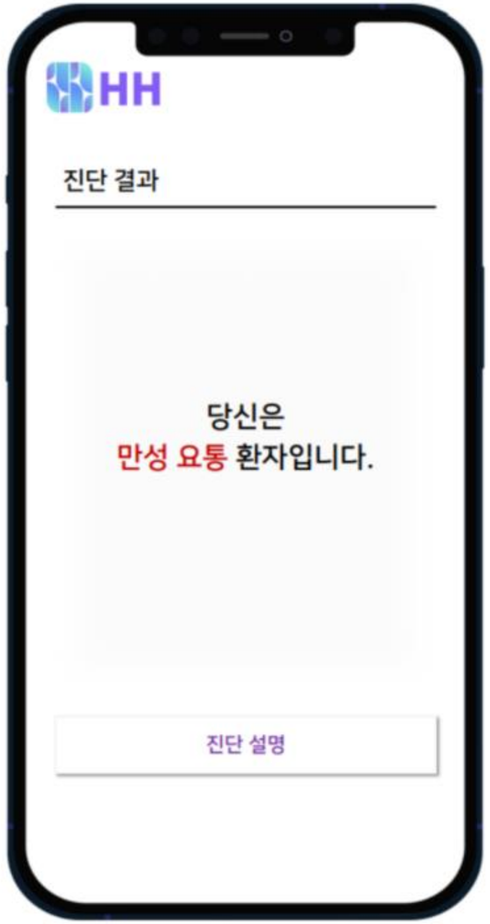


변수 기여도

변수	User
1번 자세	32%
2번 자세	46%
3번 자세	2%
4, 5번 자세	1%
6번 자세	2%
7번 자세	2%
8, 9번 자세	1%
10, 11번 자세	15%
2번 자세 변화	18%
8, 9번 자세 변화	-3%
10, 11번 자세 변화	9%
자세 변화 빈도	98회

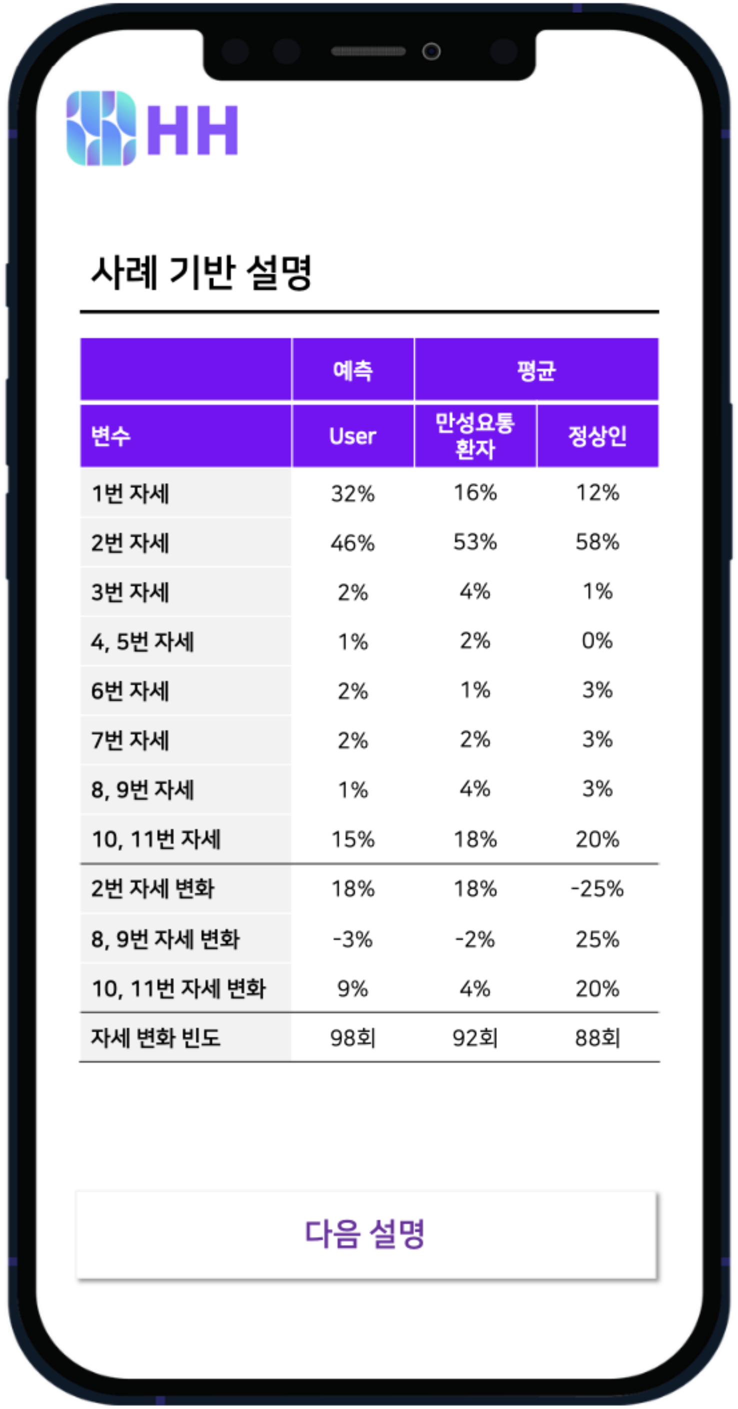
Explanation interface prototypes (diagnostic output: ‘CLBP patient’)

- Example-based explanation



진단 결과

당신은
만성 요통 환자입니다.

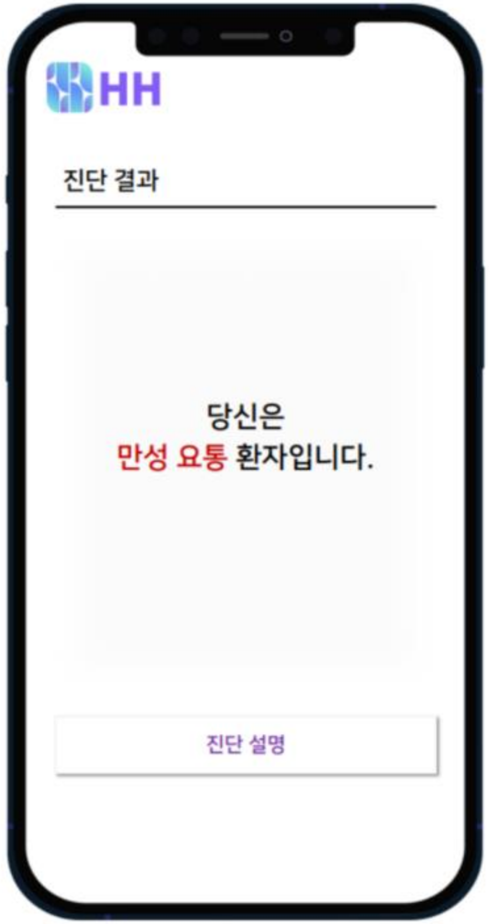


사례 기반 설명

	예측	평균	
변수	User	만성요통 환자	정상인
1번 자세	32%	16%	12%
2번 자세	46%	53%	58%
3번 자세	2%	4%	1%
4, 5번 자세	1%	2%	0%
6번 자세	2%	1%	3%
7번 자세	2%	2%	3%
8, 9번 자세	1%	4%	3%
10, 11번 자세	15%	18%	20%
2번 자세 변화	18%	18%	-25%
8, 9번 자세 변화	-3%	-2%	25%
10, 11번 자세 변화	9%	4%	20%
자세 변화 빈도	98회	92회	88회

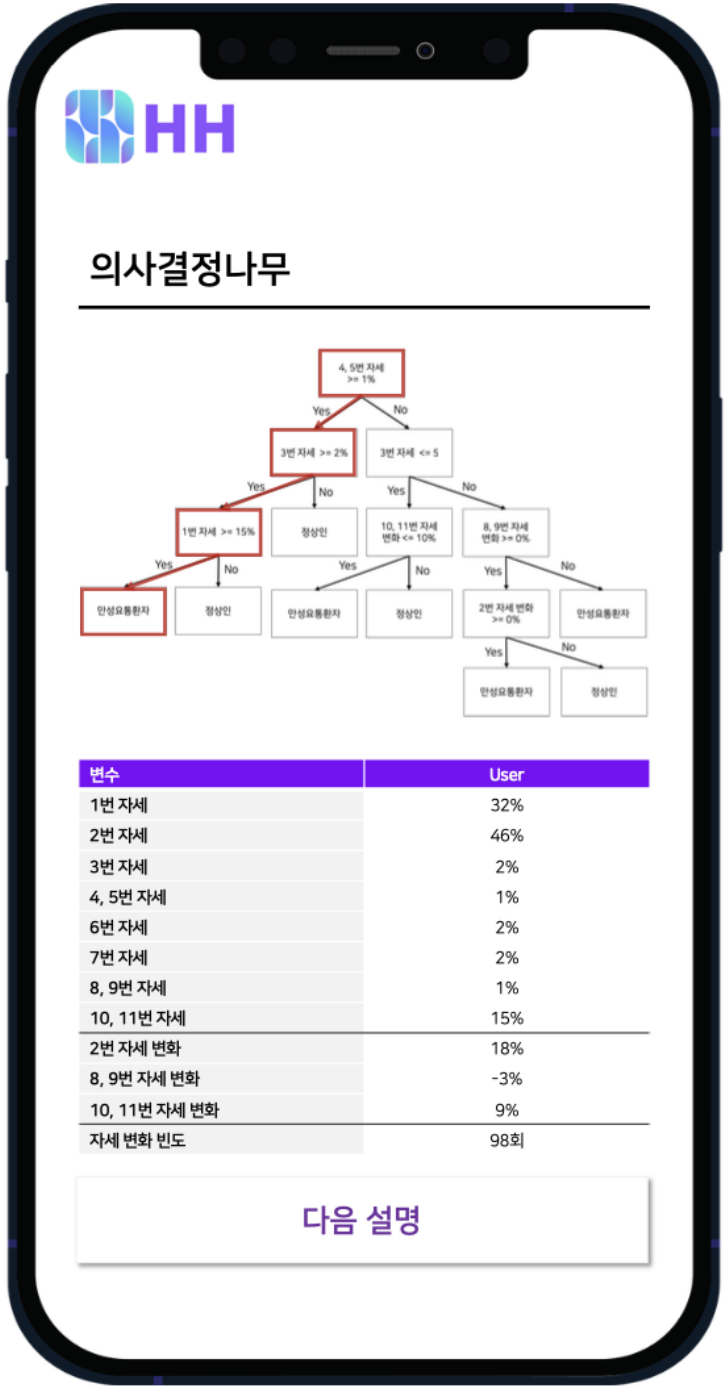
Explanation interface prototypes (diagnostic output: ‘CLBP patient’)

- Decision tree explanation



진단 결과

당신은
만성 요통 환자입니다.

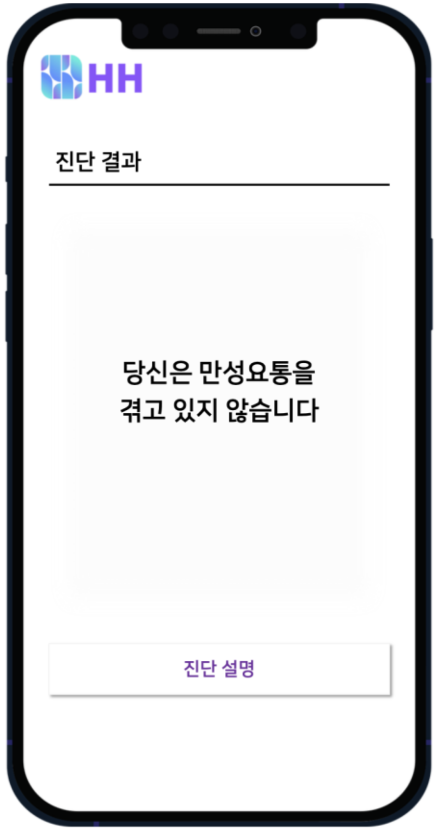


의사결정나무

변수	User
1번 자세	32%
2번 자세	46%
3번 자세	2%
4, 5번 자세	1%
6번 자세	2%
7번 자세	2%
8, 9번 자세	1%
10, 11번 자세	15%
2번 자세 변화	18%
8, 9번 자세 변화	-3%
10, 11번 자세 변화	9%
자세 변화 빈도	98회

Explanation interface prototypes (diagnostic output: ‘non-patient’)

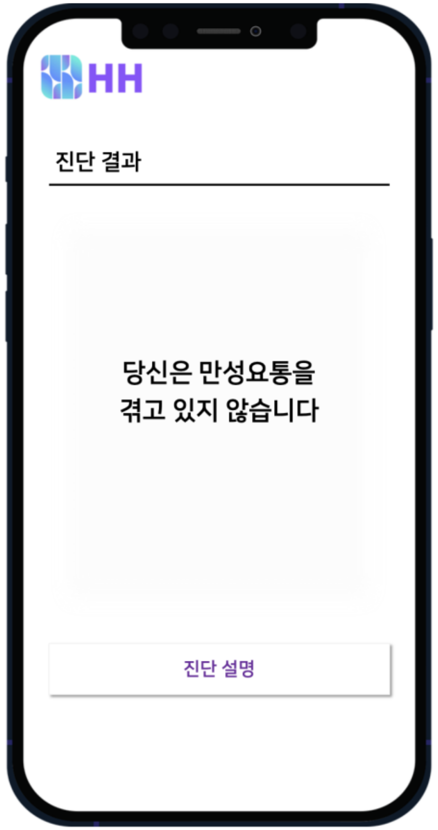
- No explanation



당신은 만성요통을
겪고 있지 않습니다

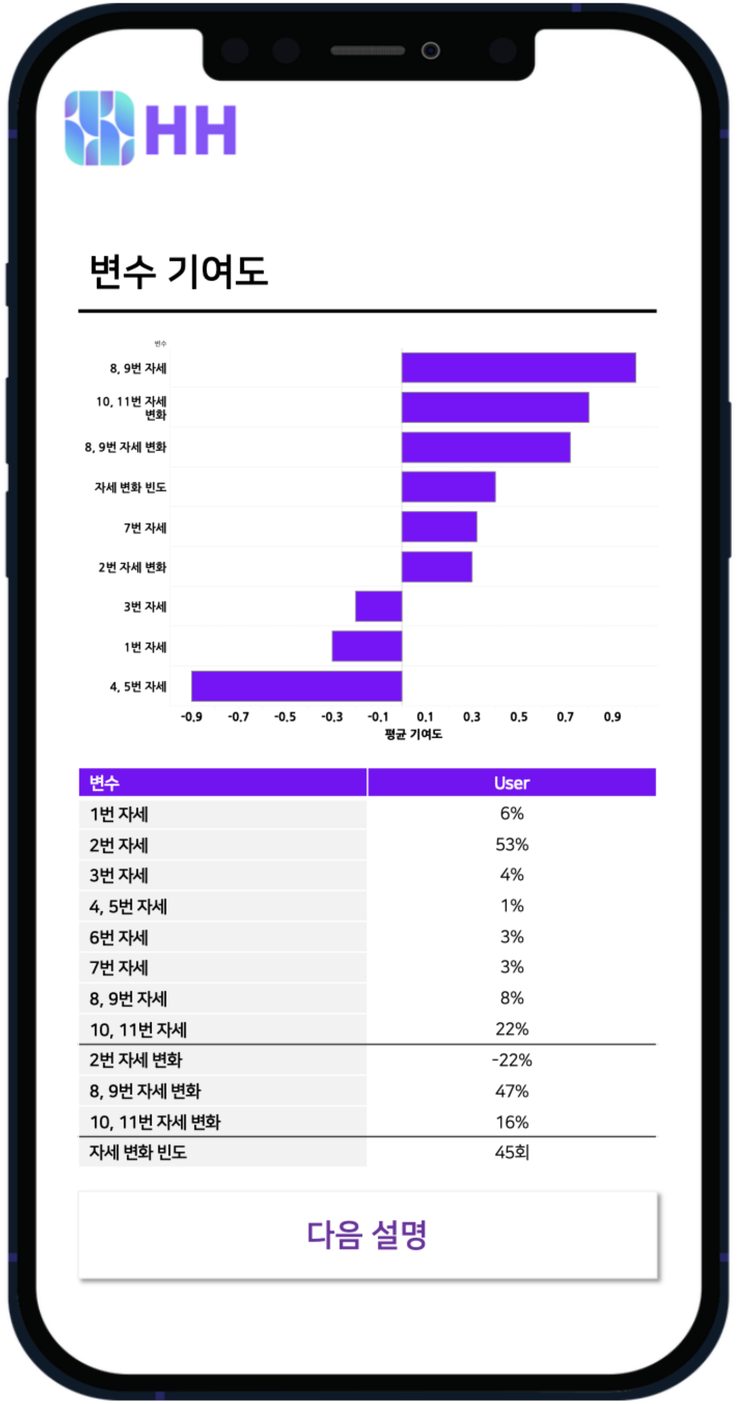
Explanation interface prototypes (diagnostic output: ‘non-patient’)

- Feature attribution explanation



진단 결과

당신은 만성요통을
겪고 있지 않습니다

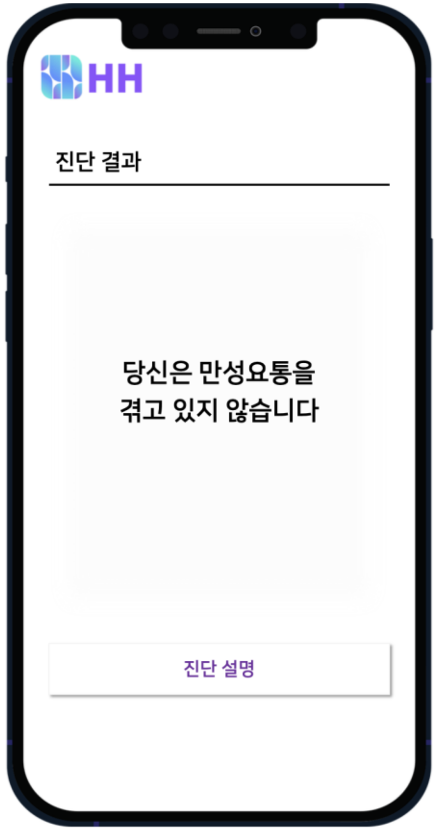


변수 기여도

변수	User
1번 자세	6%
2번 자세	53%
3번 자세	4%
4, 5번 자세	1%
6번 자세	3%
7번 자세	3%
8, 9번 자세	8%
10, 11번 자세	22%
2번 자세 변화	-22%
8, 9번 자세 변화	47%
10, 11번 자세 변화	16%
자세 변화 빈도	45회

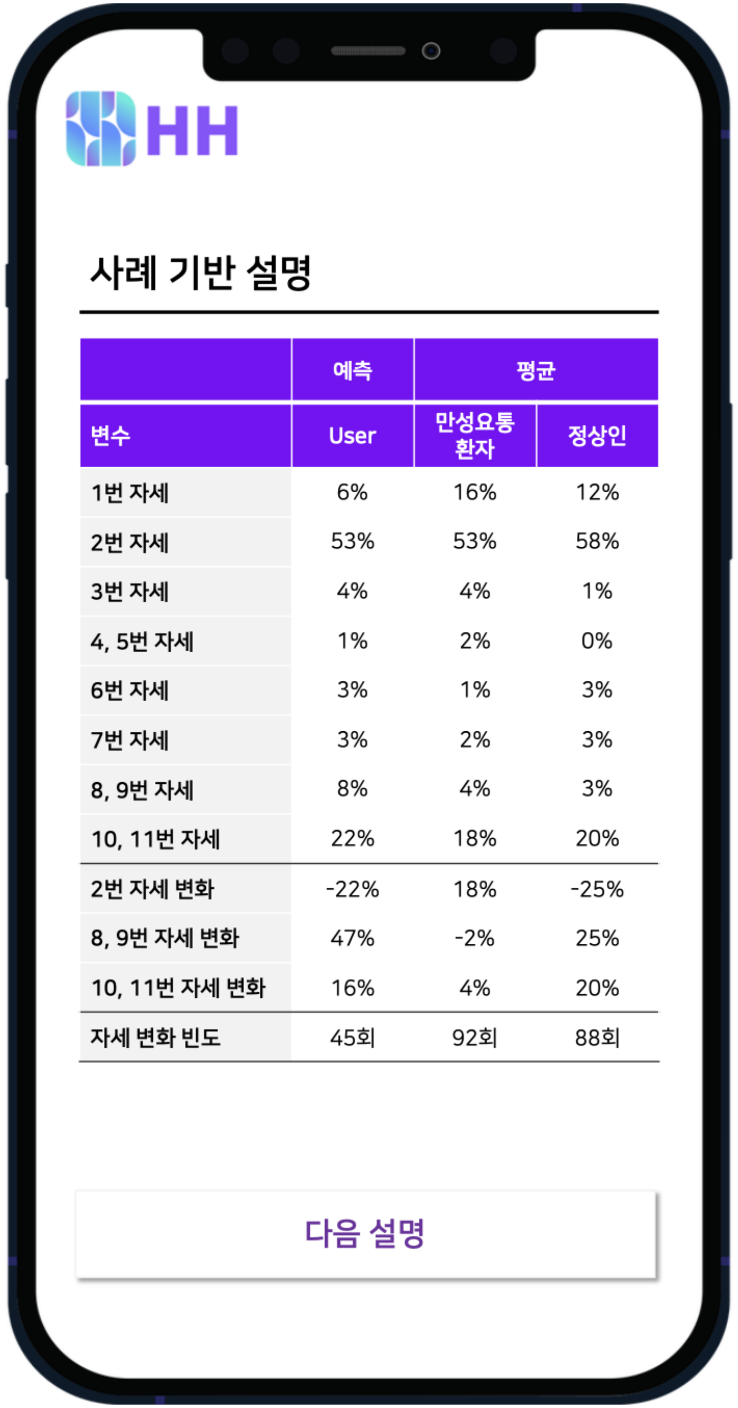
Explanation interface prototypes (diagnostic output: ‘non-patient’)

- Example-based explanation



진단 결과

당신은 만성요통을
겪고 있지 않습니다

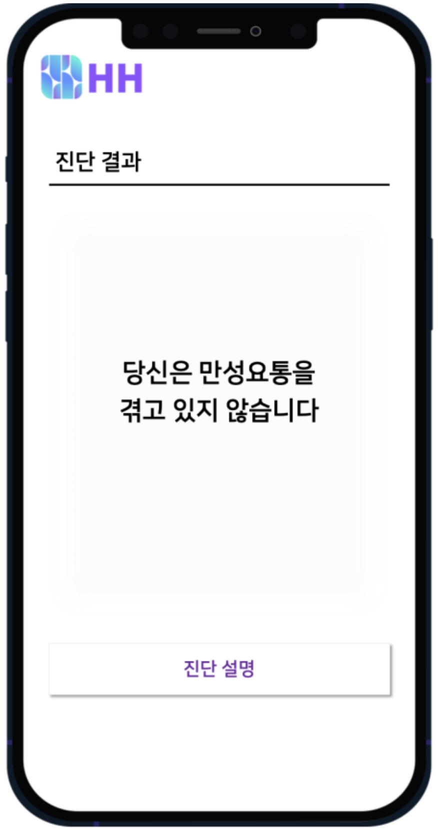


사례 기반 설명

	예측	평균	
변수	User	만성요통 환자	정상인
1번 자세	6%	16%	12%
2번 자세	53%	53%	58%
3번 자세	4%	4%	1%
4, 5번 자세	1%	2%	0%
6번 자세	3%	1%	3%
7번 자세	3%	2%	3%
8, 9번 자세	8%	4%	3%
10, 11번 자세	22%	18%	20%
2번 자세 변화	-22%	18%	-25%
8, 9번 자세 변화	47%	-2%	25%
10, 11번 자세 변화	16%	4%	20%
자세 변화 빈도	45회	92회	88회

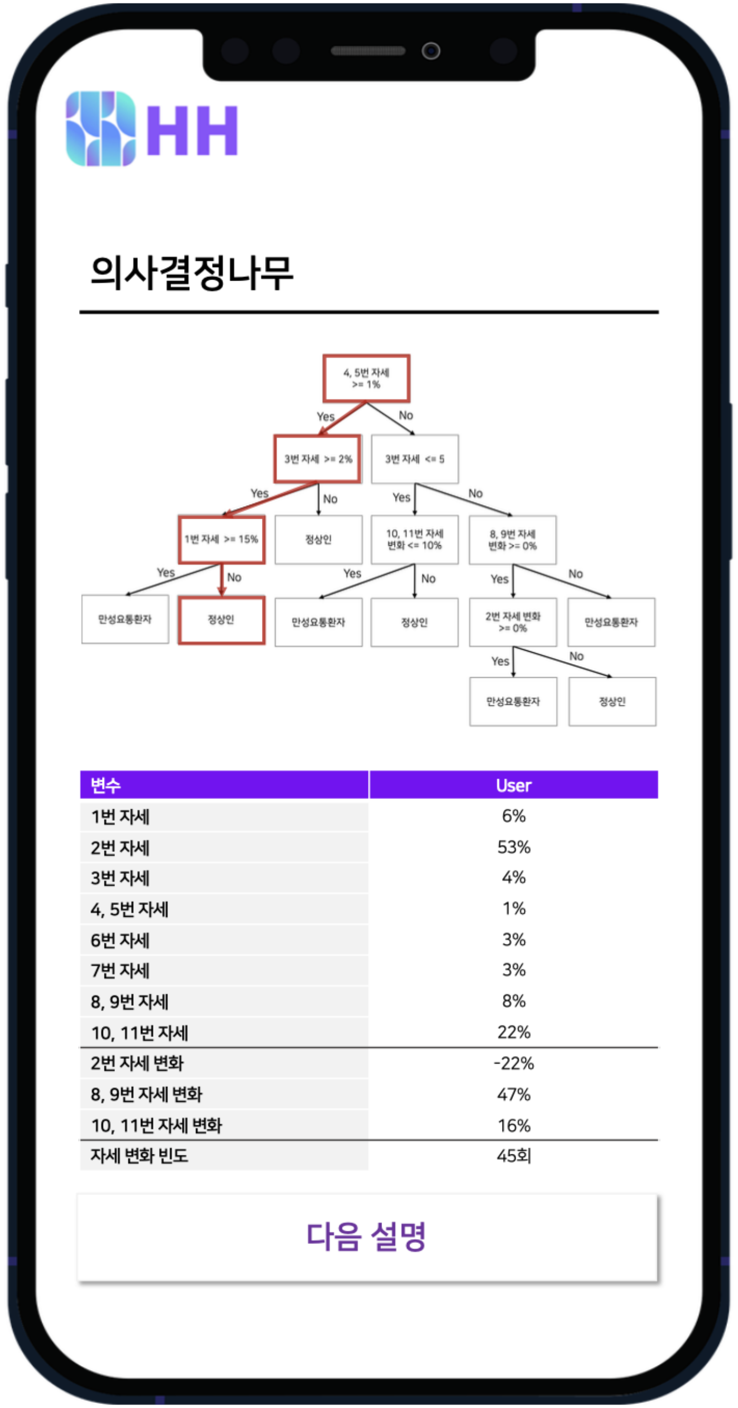
Explanation interface prototypes (diagnostic output: ‘non-patient’)

- Decision tree explanation



진단 결과

당신은 만성요통을
겪고 있지 않습니다



의사결정나무

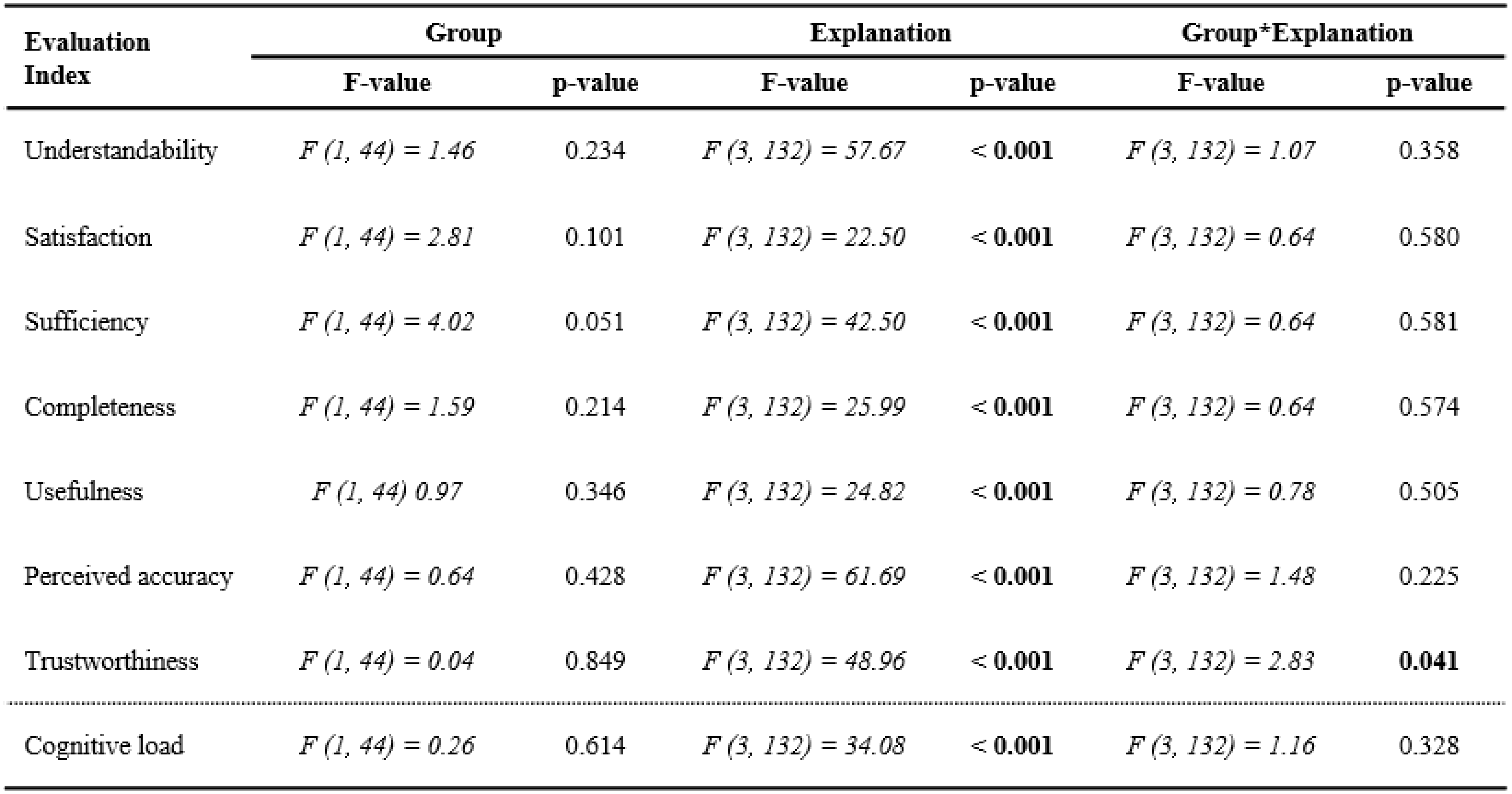
변수	User
1번 자세	6%
2번 자세	53%
3번 자세	4%
4, 5번 자세	1%
6번 자세	3%
7번 자세	3%
8, 9번 자세	8%
10, 11번 자세	22%
2번 자세 변화	-22%
8, 9번 자세 변화	47%
10, 11번 자세 변화	16%
자세 변화 빈도	45회

Experiment

- Experimental variables:
  - Independent variables:
    - Explanation type: no explanation, feature attribution, example-based and decision tree
    - Group (diagnostic output): CLBP patient, non-patient
  - Dependent variables:
    - 7 UX measures
    - Cognitive load rating
- Data analysis:
  - ANOVAs
  - Post-hoc pairwise comparisons

Results

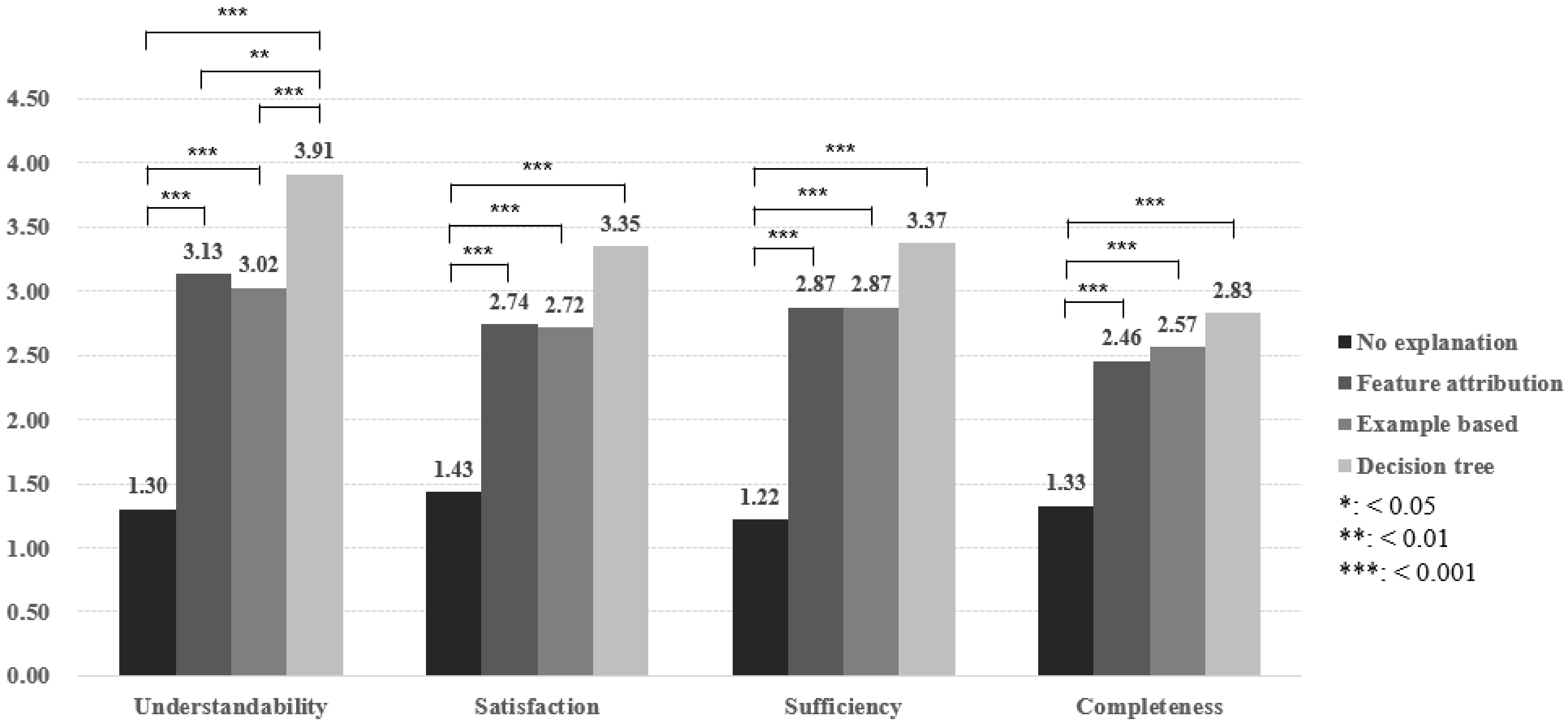
- ANOVA results



Evaluation Index	**Group**		**Explanation**		**Group*Explanation**	
	F-value	**p-value**	**F-value**	**p-value**	**F-value**	**p-value**
Understandability	*F (1, 44) = 1.46*	0.234	*F (3, 132) = 57.67*	< **0.001**	*F (3, 132) = 1.07*	0.358
Satisfaction	*F (1, 44) = 2.81*	0.101	*F (3, 132) = 22.50*	< **0.001**	*F (3, 132) = 0.64*	0.580
Sufficiency	*F (1, 44) = 4.02*	0.051	*F (3, 132) = 42.50*	< **0.001**	*F (3, 132) = 0.64*	0.581
Completeness	*F (1, 44) = 1.59*	0.214	*F (3, 132) = 25.99*	< **0.001**	*F (3, 132) = 0.64*	0.574
Usefulness	*F (1, 44) 0.97*	0.346	*F (3, 132) = 24.82*	< **0.001**	*F (3, 132) = 0.78*	0.505
Perceived accuracy	*F (1, 44) = 0.64*	0.428	*F (3, 132) = 61.69*	< **0.001**	*F (3, 132) = 1.48*	0.225
Trustworthiness	*F (1, 44) = 0.04*	0.849	*F (3, 132) = 48.96*	< **0.001**	*F (3, 132) = 2.83*	**0.041**
Cognitive load	*F (1, 44) = 0.26*	0.614	*F (3, 132) = 34.08*	< **0.001**	*F (3, 132) = 1.16*	0.328

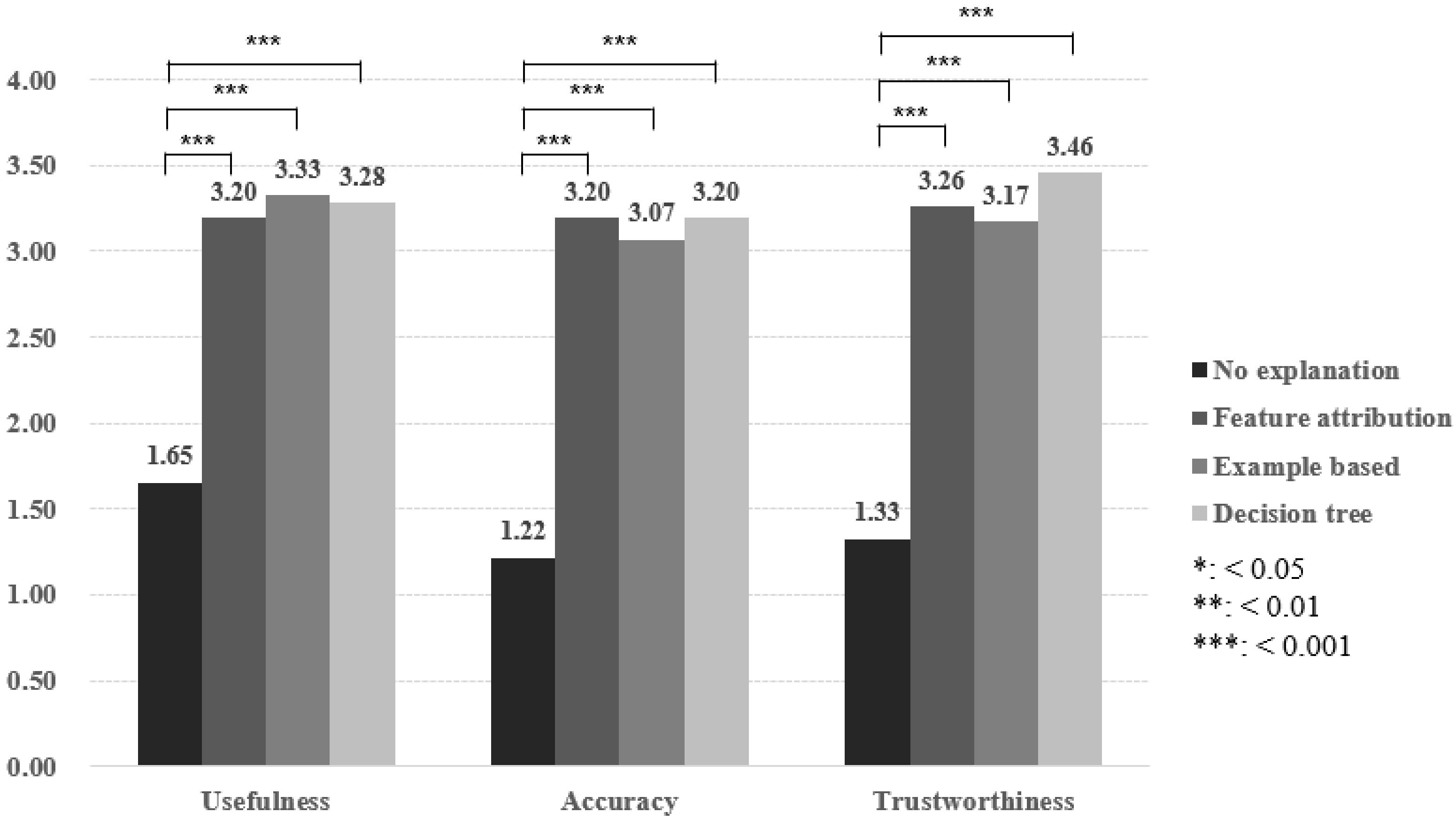
Results

- ‘Explanation type’ main effects: UX measures



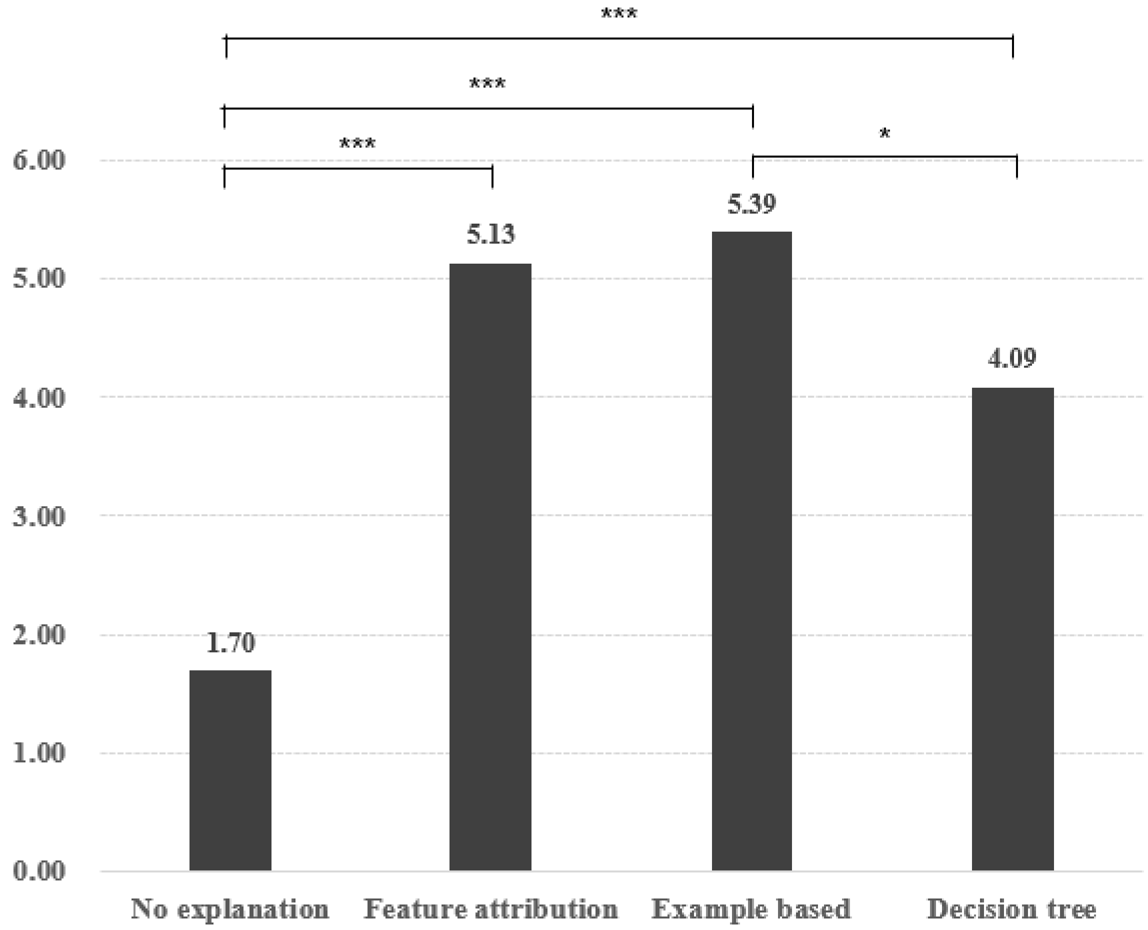
Results

- ‘Explanation type’ main effects: UX measures



Results

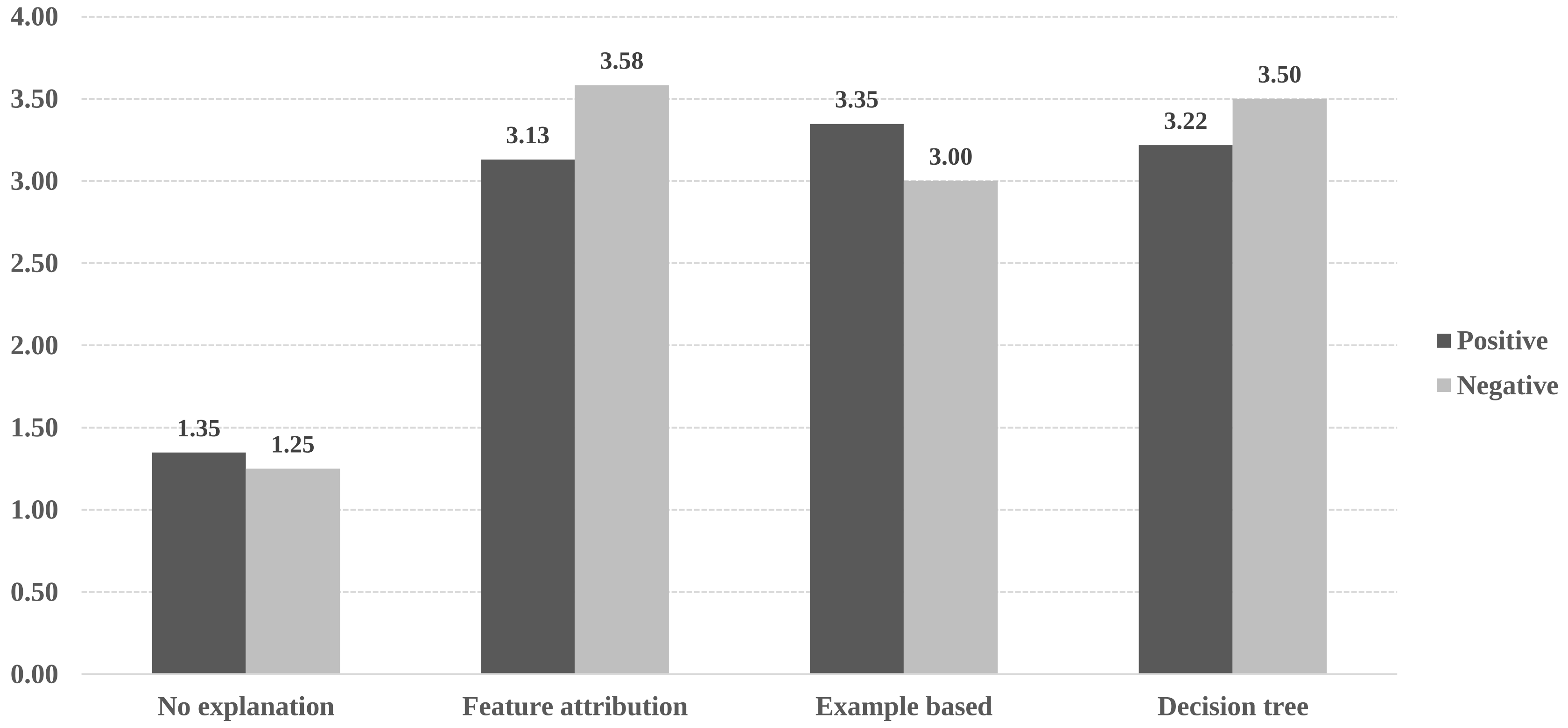
- ‘Explanation type’ main effects: cognitive load



*: < 0.05
**: < 0.01
***: < 0.001

Results

- ‘Explanation type x group (diagnostic output)’ interaction effect: perceived trustworthiness (trust)



Discussion

- Compared with no explanation, the three explanation types resulted in better user responses. This indicates that the existing XAI methods are useful.
- Overall, decision tree explanation appeared more advantageous compared with the other alternatives.
    - Decision tree explanation was found to be significantly better than the other alternatives in terms of understandability.
    - Decision tree explanation also resulted in a significantly lower mean cognitive load score than example-based explanation.

Discussion

- Decision tree explanation was found to be significantly better than the other alternatives in terms of understandability. This may be because:
  - People are already familiar with the use of “IF ~ THEN ~” rules in describing a decision process.
  - A set of “IF ~ THEN~” rules (decision tree explanation) fully and directly describes how a decision is made. On the other hand, feature attribution and example cases do not. They are indirect at best in describing a decision process.
  - Understanding the notions, feature importance and example cases, requires some mental models associated with AI/ML methods.
  - When using decision tree explanation, one only needs to process one “IF ~ THEN ~” rule at a time – a decision process can be broken down into a series of easy-to-process rules. On the other hand, feature attribution explanation requires integrating multiple feature importance values. Example-based explanation also requires comparing the case of interest with each of the examples presented in the multi-item vector representation. Such mental integration is demanding and would compromise understandability.

Discussion

- Decision tree explanation also resulted in a significantly lower mean cognitive load score than example-based explanation. Again, this could be explained on the basis of differences in human information processing requirements:
  - When using decision tree explanation, one only needs to process one “IF ~ THEN ~” rule at a time – a decision process can be broken down into a series of easy-to-process rules. Example-based explanation also requires comparing the case of interest with each of the examples presented in the multi-item vector representation. Such mental integration is demanding and thus increases cognitive loads.

Discussion

- The ‘explanation type x group (diagnostic output)’ interaction effect on perceived trustworthiness (trust) was statistically significant, indicating that the utility of a particular explanation type may change according to what the machine diagnosis is. However, the interaction effect was rather small.
  - This finding seems to suggest that the utility of a particular explanation type/method would change according to a machine decision (diagnosis) and its implications. This warrants further investigations.

Discussion

- Whilst the explanation types considered were found to be useful compared with ‘no explanation,’ their mean ratings were not great. For the UX measures, none of the mean values were greater than 4 on the 5-point scale. Also, the mean cognitive load scores were greater than 4 on the 9-point scale. There is room for improvement.
  - The explanations based on the existing XAI methods may not address the real information needs of the users in the particular context of a smart-chair based CLBP recognition system.
  - Pre-determining user information needs based on existing XAI methods may not be effective. A better approach might be to discover user explanation needs specific to each particular AI application context through some systematic contextual inquiry and analysis.

Conclusions

- Overall, the existing XAI methods were found to be useful. Especially, decision tree based explanation seemed better than feature importance and example-based explanations.
- Despite their utility, however, the existing XAI methods and the information they provide may not fully address the users' information needs.
- Methods for identifying the user information needs specific to each particular AI application (e.g., telediagnosis) domain are needed.

References

- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. Journal of educational psychology, 84(4), 429.
- Jeong, H., & Park, W. (2020). Developing and evaluating a mixed sensor smart chair system for real-time posture classification: Combining pressure and distance sensors. IEEE Journal of Biomedical and Health Informatics, 25(5), 1805-1813.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.
- Kilgore, R., & Voshell, M. (2014). Increasing the transparency of unmanned systems: Applications of ecological interface design. In Virtual, Augmented and Mixed Reality. Applications of Virtual and Augmented Reality: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 6 (pp. 378-389). Springer International Publishing.
- Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the need for explainable artificial intelligence (xAI). arXiv preprint arXiv:2012.01007.