Human-AI Collaboration in Healthcare

Albert Tanubrata
Computer Science Department, Faculty of Computer Science
Bina Nusantara University
Jakarta, Indonesia
albert.tanubrata001@binus.ac.id

Bernardus Ignasio
Computer Science Department, Faculty of Computer Science
Bina Nusantara University
Jakarta, Indonesia
bernardus.ignasio@binus.ac.id

Eko Setyo Purwanto, S.Pd., M.Kom.
D6659
Bina Nusantara University
Jakarta, Indonesia
eko.purwanto@binus.ac.id

Muhamad Keenan Ario, S.Kom, M.Kom.
D6664
Bina Nusantara University
Jakarta, Indonesia
muhamad.ario@binus.ac.id

***Abstract*—This research uses pre-processing techniques, specifically stemming and lemmatizing, to investigate the performance of various machine learning models in healthcare applications. There are several models we compare, including Random Forest, SVC, Logistic Regression, and several ensemble methods, across key metrics such as accuracy, precision, and F1 score. The results show that pre-processing techniques can slightly reduce the performance of models such as Random Forest and Voting Classifier maintaining high performance, thereby demonstrating their robustness.**

Keywords—Human-AI collaboration, healthcare, machine learning, text preprocessing, stemming, lemmatizing, model evaluation, accuracy, precision, F1-score, Random Forest, ensemble methods.

I. Introduction

In this modern era, Artificial Intelligence(AI) is increasingly advancing and affecting humans in many factors. With the ability of learning data, identifying patterns, providing knowledge, and making decisions, after years of developments, finally AI has capable to adapt human's intelligence and proven worthy to be developed deeper in many sectors such as education, business, economic, law, and also medical in order to make human's life easier and achieving incapable things. However, This breakthrough doesn't mean that AI can fully replace and take over human's role in this life. As an artificial innovation, There are also errors and confusion in AI, no matter how small it is.

One of the major sectors that require AI implementation is the medical sector. It had been stated that there is a lack of medical professionals and inequality betweens number of medical professionals and patients [1]. The World Health Organization (WHO) even predicted a shortfall of 10 million of medical professionals especially in lower and lower middle income countries. This scarcity of human resources leads to an overwhelmed situation in the medical sector, such as when the COVID-19 striked. These circumstances cause reduction of medical quality,

possibly misdiagnosed cases, and also overworked medical professionals[1].

This condition provides opportunities for AI to improve quality of healthcare and workload efficiency, considering AI's track record on decision making capability on clinical tasks such as diagnosing cancer in histopathology, detecting diabetic retinopathy, or evaluating X-Ray for disease identification such as pneumonia[2]. However, this record doesn't mean that AI can replace medical professionals in some tasks. From previous study stated that in many parts, Collaboration between humans and AI is more efficient and effective, better than replacing humans with AI completely[3]. This concludes that instead of depending completely on AI, applying Human-AI Collaboration is preferably done in the medical field especially in diagnosing and identifying diseases.

There are still plenty of issues in implementing Human-AI Collaboration in Healthcare, specifically in decision making about diagnosis and disease identification. From the previous research, the problem came from both inner, such as possibilities of unsuccessful prediction, lack of high-quality data, inaccurate output, and human outperformed [4], and also outer such as over-reliance, under-reliance, and opacity of judgment's reliability on AI diagnosis[5]. These issues cause misdiagnosis, uncredible medical professionals, and also endanger patients life which are major problems in further development of Human-AI Collaboration in Healthcare Diagnosis. This turns the good intention to improve medical quality into worsening the medical quality.

The purpose of this paper is to assist medical sectors and improve medical quality by analyzing the effective development of Human-AI Collaboration In Healthcare Diagnosis in techniques such as Machine Learning (ML), and DL (Deep Learning). Our research is expected to reduce the workload of medical professionals, cost and also time of medication with Human-AI Collaboration such as supervised AI diagnosis by medical professionals. The first section from this paper is conducting literature review based on related papers about Human-AI Collaboration, then analyzing the techniques in which can develop more accurate medical diagnosis and the way of collaboration with medical professionals, and last analyzing ways to prevent outer issues of Human-AI Collaboration.

II. Related Research

Collaboration between humans and artificial intelligence (AI) in health care has become an increasingly interesting topic in this modern era. The use of AI in the health sector provides many benefits, including improving diagnostic accuracy, operational efficiency and medical data management. This implementation of AI in health care also faces various challenges, such as ethical issues, data privacy, and acceptance from medical personnel. Therefore it is important to explore how collaboration between humans and AI can be optimized to achieve the best outcomes in healthcare.

Various studies have been conducted to explore the potential of human-AI collaboration in the health sector. One method that is often used is machine learning (ML) to analyze medical data. Machine learning (ML) algorithms such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been proven effective in diagnosing diseases based on medical images and patient data[5].

Tumors in the brain are abnormal growths of cells in the brain that can be malignant (cancerous) or non-malignant (non-cancerous). Tumors in the brain can originate from the brain cells themselves (primary tumors) or can spread from other parts of the body (secondary or metastatic tumors).

Human-AI collaboration in the health sector can be applied to tumors that often occur in the medical world. Tumors in the brain can interfere with brain function, causing symptoms that vary depending on the location, size and type of tumor. Symptoms that may appear include headaches, nausea and vomiting, changes in vision or hearing, weakness or tingling on one side of the body, changes in behavior or personality, and problems with coordination and balance.

Recent years have witnessed a surge in research efforts aimed at leveraging AI technology to improve brain tumor detection[18]. Various studies have explored the efficacy of deep learning algorithms, such as convolutional neural networks (CNN), in analyzing medical imaging data,

including magnetic resonance (MRI) scans and computed tomography (CT)[7].

Despite promising advances, the integration of AI into brain tumor detection is not without challenges. One of the main concerns is the interpretability of results generated by AI, especially in complex medical contexts[8]. Ensuring the transparency and reliability of AI algorithms remains a pressing issue, as incorrect interpretations can have serious implications for patient care.

Additionally, the scarcity of annotated medical imaging data is a significant obstacle to the development and validation of AI models for brain tumor detection[9]. Limited access to high-quality datasets hinders the robustness and generalisability of AI algorithms, highlighting the need for collaborative efforts to address these data gaps.

Based on various previous studies, it can be concluded that collaboration between humans and AI has great potential to improve the quality of health services. To achieve effective collaboration, a holistic approach is needed, including aspects of technology, eth

III. Methodology

To identify the required form of collaboration between professional and AI, a systematic literature review is conducted. To collect relevant papers, three major keywords such as "Human-AI Collaboration in Healthcare", "AI Role in Medical Field ", "AI in Healthcare Diagnosis" were used on ResearchGate, Scopus, and Google Scholar, resulting in 30 possibly related papers from 2018-2023, to be sorted and reviewed to obtain 2 suitable paper to identify the required form.

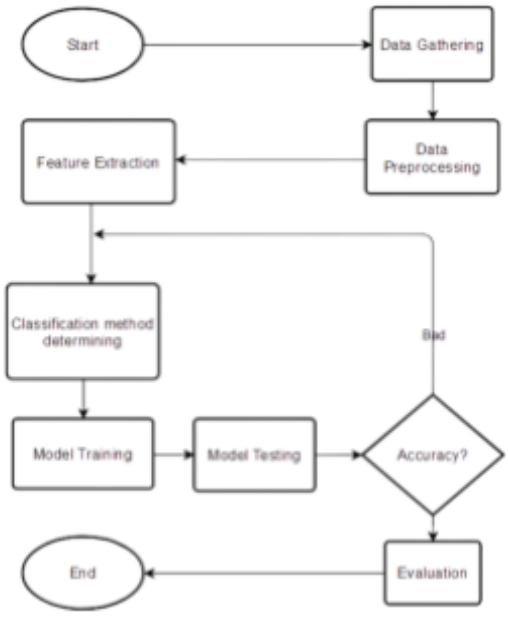


Figure 3.1 : Experiment Flowchart

To analyze effective AI techniques to support collaboration between Healthcare Professionals and AI, an experiment on comparing AI techniques is conducted (**Figure 3.1**). This requires disease datasets, and training data with multiple AI algorithms.

To collect datasets, Disease datasets were searched through Kaggle with keywords such as "Disease Diagnosis", "Disease Symptoms", and "Healthcare Diagnosis" resulting in 2 major disease datasets from 2020 and 2023 with 9.12 (**Figure 3.2)** and 10.0 **(Figure 3.3**) usability score

	Unnamed: 0	label	text
0	0	Psoriasis	I have been experiencing a skin rash on my arm...
1	1	Psoriasis	My skin has been peeling, especially on my kne...
2	2	Psoriasis	I have been experiencing joint pain in my fing...
3	3	Psoriasis	There is a silver like dusting on my skin, esp...
4	4	Psoriasis	My nails have small dents or pits in them, and...

Figure 3.2 : Dataset 1

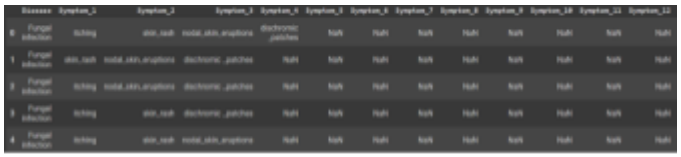


Figure 3.3 : Dataset 2

which contains 46 different diseases and 6120 rows of data in total.

Collected data should be prepared before being used in training. This requires preprocessing the data by combining the datasets, equalizing and synchronizing the datasets, and applying Natural Language Processing method to preprocessing text. Combining, equalizing and synchronizing the datasets is necessary remembering that the data used is separated into 2 datasets, and each of them has a different format. Tokenizing text into words, removing the english stopwords, lowering and filtering punctuations are NLP methods to preprocessing text, this is usable to clean the data, and improve the accuracy of prediction of each technique.

Processed data should be evaluated to get the extracted features. This requires exploring relevant features to predict diseases and encoding the features. To explore features, exploratory data analysis is conducted by exploring the distribution

of each disease, and finding the common symptoms of the diseases. Explored features then get encoded into suitable form to predict diagnosis with TF-IDF vectorizer for the symptoms and label encoder for diseases. Lastly, encoded features separated into training and testing size.

Classification algorithms for classifying text such as Machine Learning approaches like Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, or even Deep Learning approaches such as Recurrent Neural Network (RNN), and BERT are used to fit the training data, and then predict the testing data. Analyzing accuracy of each algorithm to see and determine the best algorithms to diagnose diseases. Evaluation metrics are then used for evaluating the algorithm with methods such as F1 score, Confusion Matrix, Accuracy, Precision, and Recall.

IV. Result & Discussion

4.1. *Usage of Health Diagnostic AI accordingly to Medical Professional*

Overall, researchers have started to utilize artificial intelligence (AI) in diagnosing health by taking various approaches to produce optimal results. AI offers new avenues for advancing technology to enhance accuracy, speed, and efficiency in diagnosing various health conditions. Similar to systematic literature reviews, numerous researchers are already integrating AI into disease diagnosis, utilizing decision-making processes and AI tools to streamline the diagnostic process. In addition to aiding diagnosis, AI in health diagnosis also facilitates personalized treatment plans tailored to individual patients, optimizing therapeutic interventions and ultimately improving patient outcomes. The integration of AI enables the analysis of vast amounts of patient data, leading to the discovery of novel biomarkers and disease patterns that may have previously gone unnoticed. This data-driven approach not only enhances diagnostic accuracy but also contributes to a deeper understanding of disease mechanisms, paving the way for more effective treatment strategies in the future.

4.2. *Effective AI techniques in Medical Diagnosis*

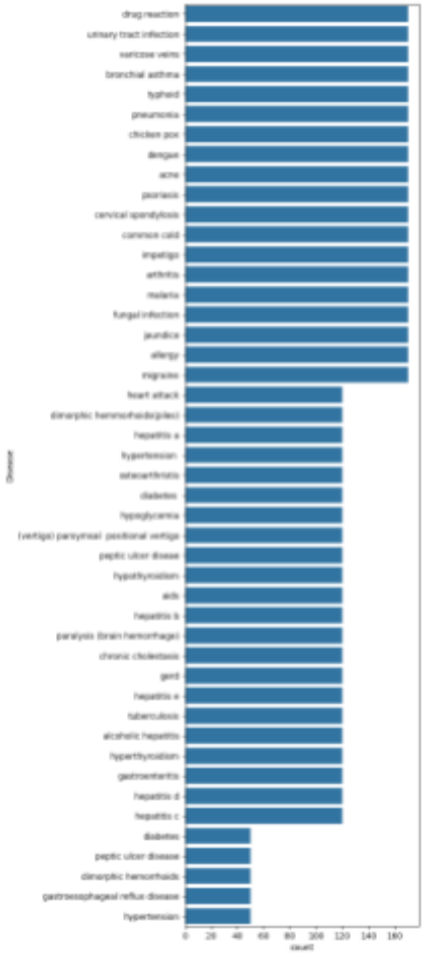
In order to reach the goal to find most effective AI techniques to diagnose diseases based on text symptoms, the labeled datasets with different

	Disease	text
2094	drug reaction	itching, skin rash, stomach pain, burning mict...
1612	urinary tract infection	bladder discomfort, foul smell of urine, cont...
5858	urinary tract infection	I need to relieve myself regularly, but I can'...
4294	varicose veins	fatigue, cramps, bruising, obesity, swollen l...
2571	heart attack	vomiting, breathlessness, sweating, chest pain
...	...	...
3772	fungal infection	itching, skin rash, nodal skin eruptions, disc...
5191	dengue	I'm experiencing extreme body pain, headache a...
5226	fungal infection	My skin has been itching a lot and developing ...
5390	dimorphic hemorrhoids	Since I've been constipated, using the restroo...
860	drug reaction	itching, skin rash, stomach pain, burning mict...

Figure 4.1 : Joined Dataset

formats are joined to enrich datasets with different types of text. As seen in **figure 4.1**. There are differences between text formats, with the aim to build more accurate models. The joined dataset has

Figure 4.2 : Data Distribution



46 Different diseases with even distribution of every disease such as seen in **figure 4.2**. The gap between the number of diseases is approximately around dozens. There are also commons symptoms in the dataset that mostly mentioned such as high fever, yellowish skin, abdominal pain, etc (**figure 4.3**)

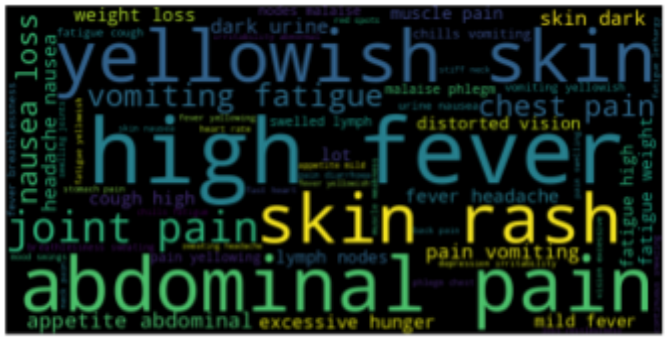


Figure 4.3 : Common Symptoms

Preprocessing the data then is done with only tokenizing and removing stopwords from the data, then encoded and vectorized the data. Cleaned processed data then being fitted into several algorithms for text classification and the effectiveness is calculated with evaluation metrics. Based on the evaluation metrics analysis, implementation of stemming and lemmatization, also different approaches of some classifier models are reducing the number of evaluation metrics. The comparison can be seen in Table 1 and Table 2.

Table 4.1 : Without Stemming & Lemmatizing

No	Evaluation Metrics			
	Model	Accuracy	Precision	F1-Score
1	Multinomial NB	96,98%	96,98 %	96,47%
2	Bernoulli NB	95,67%	95,67 %	94,91%
3	Gaussian NB	98,61%	98,61 %	98,60%
4	Logistic Regression	99,43%	99,43 %	99,14%
5	SVC	99,67%	99,67 %	99,67%
6	XGBOOST	98,61%	98,61 %	98,60%
7	Bagging Classifier	98,69%	98,69 %	98,68%
8	Voting Classifier ‘Soft’	99,84%	99,84 %	99,84%
9	Voting Classifier ‘Hard’	99,51%	99,51 %	99,50%
10	Random Forest Classifier (20 n_estimator)	100%	100%	100%
11	BERT (1 Epochs)	90,83%	90,52 %	-
12	RNN (10 Epochs)	93,71%	93,71 %	-

Table 4.2: With Stemming & Lemmatizing

No	Evaluation Metrics			
	Model	Accuracy	Precision	F1-Score
1	Multinomial NB	96,49%	96,49 %	95,84%
2	Bernoulli NB	96,98%	96,98 %	96,69%
3	Gaussian NB	98,53%	98,53 %	98,52%
4	Logistic Regression	99,35%	99,35 %	99,33%
5	SVC	99,51%	99,51 %	99,51%
6	XGBOOST	98,12%	98,12	98,10%

			%	
7	Bagging Classifier (20 n_estimator)	97,88%	97,88 %	97,85%
8	Voting Classifier ‘Soft’	99,75%	99,75 %	99,75%
9	Voting Classifier ‘Hard’	99,35%	99,35 %	99,34%
9	Random Forest Classifier (20 n_estimator)	100%	100%	100%
10	BERT	90,26%	90,25 %	
11	RNN	92,48%	92,48 %	-

The percentage differences can be seen clearly. In this case without lemmatization and stemming, most of the models perform better, although in some cases like Bernoulli Naive Bayes, using lemmatization and stemming makes the performance better. This might be the effect of lemmatization and stemming. It sometimes makes the text unclear and erases some alphabets in words.

From **table 1**, performances of each model are great, but with the same approach and quick duration noted by the n_estimators is only 20, ensemble learning models outperform the conventional models with a difference of about 2-3%. Besides that, Deep learning approaches also perform great on the evaluation metrics, but require more time and high specification devices (bigger epochs, more accurate), but with neural network approaches, this model might outperform and be more detailed in bigger tasks.

Overall, in this classification task, the effectiveness difference can be seen by the evaluation metrics, and duration of training models. The utilization of the models still relies on medical professionals' conditions and needs, but based on this comparison, the ensemble and deep learning method are more recommended in general diagnosing tasks.

V. Conclusion

In this study we explore the impact of pre-processing techniques, specifically stemming and lemmatizing, on the performance of various machine learning models in healthcare applications. The comparisons were made on several evaluation metrics including accuracy, precision, and F1 score.

In **table 4.1,** it can be seen that the Random Forest Classifier (20 n_estimator) achieved the highest performance with accuracy, precision and F1 score of 100%. Then the 'Soft' Voting classifier followed with a performance of 99.84% across all metrics. And SVC also shows high performance with accuracy, precision and F1 score of 99.67%.

In **table 4.2,** it can be seen that the Random Forest Classifier (20_n estimator) again achieved the highest performance with accuracy, precision and F1 score of 100%. Then the 'Soft' voting classifier has slightly reduced performance compared to without preprocessing, with accuracy, precision and F1 score of 99.75%. And SVC also shows a slight performance drop with all metrics at 99.51%.

Overall it can be said that the stemming and lemmatization pre-processing techniques show mixed results. For most models, there is a slight decrease in performance when this technique is applied. For example, the accuracy of the Logistic Regression model decreased from 99.43% to 99.35%, and the F1 score from 99.14% to 99.33%. Likewise, the performance of the Bagging Classifier drops more markedly from 98.69% to 97.88%.

Despite these variations, the Random Forest Classifier consistently maintains the highest performance in both scenarios, demonstrating its robustness and reliability in handling textual data in healthcare applications. In addition, ensemble methods such as 'Soft' and 'Hard' Voting Classifiers also perform very well, thus strengthening the effectiveness of combining multiple models to improve prediction accuracy.

So the conclusion obtained after conducting experiments in this research is that although stemming and lemmatization can be useful in reducing computational complexity, their impact on model performance can vary. Careful consideration is required when deciding to implement these steps in preprocessing, as their effectiveness may depend on the specific characteristics of the data set and the machine learning model chosen.

References

[1] Y. Lai, A. Kankanhalli, and D. Ong, *Human-AI Collaboration in Healthcare: A Review and Research Agenda*. 2021. doi: 10.24251/HICSS.2021.046.

[2] P. Hemmer, M. Schemmer, L. Riefle, N. Rosellen, M. Vössing, and N. Kühl, *Factors that Influence the Adoption of Human-AI Collaboration in Clinical Decision-Making*. 2022. doi: 10.48550/arXiv.2204.09082.

[3] A. Sharma, I. Lin, A. Miner, D. Atkins, and T. Althoff, "Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support," *Nat. Mach. Intell.*, vol. 5, pp. 1–12, Jan. 2023, doi: 10.1038/s42256-022-00593-2.

[4] M. Maadi, H. Akbarzadeh Khorshidi, and U. Aickelin, "A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applications," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 4, p. 2121, Feb. 2021, doi: 10.3390/ijerph18042121.

[5] C. Reverberi *et al.*, "Experimental evidence of effective human-AI collaboration in medical decision-making," *Sci. Rep.*, vol. 12, no. 1, p. 14952, Sep. 2022, doi: 10.1038/s41598-022-18751-2.

[6] S. Y. Park *et al.*, "Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare," in *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, in CSCW '19 Companion. New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 506–510. doi: 10.1145/3311957.3359433.

[7] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, "'Hello AI': Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019, doi: 10.1145/3359206.

[8] F. Cabitza *et al.*, "Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis," *Artif. Intell. Med.*, vol. 138, p. 102506, Apr. 2023, doi: 10.1016/j.artmed.2023.102506.

[9] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, Jan. 2022, doi: 10.1038/s41591-021-01614-0.

[10] M. T. Sqalli, D. Al-Thani, M. Qaraqe, and L. Fernandez-Luque, "Perspectives on Human-AI Interaction Applied to Health and Wellness Management: Between Milestones and Hurdles," pp. 41–51, 2021, doi: 10.1007/978-3-030-67303-1_4.

[11] T. Karaköse, M. Demirkol, H. Köse, and R. Yirci, "A Conversation with ChatGPT about the Impact of the COVID-19 Pandemic on Education: Comparative Review Based on Human-AI Collaboration," *Educ. Process Int. J.*, vol. 12, pp. 7–25, Jul. 2023, doi: 10.22521/edupij.2023.123.1.

[12] Q. Jiang, Y. Zhang, and W. Pian, "Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic," *Inf. Process. Manag.*, vol. 59, no. 6, p. 103074, Nov. 2022, doi: 10.1016/j.ipm.2022.103074.

[13] R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, "'An Ideal Human': Expectations of AI Teammates in Human-AI Teaming," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3, pp. 1–25, Jan. 2021, doi: 10.1145/3432945.

[14] C. Rastogi, M. Tulio Ribeiro, N. King, H. Nori, and S. Amershi, "Supporting Human-AI Collaboration in Auditing LLMs with LLMs," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, Montr\'{e}al QC Canada: ACM, Aug. 2023, pp. 913–926. doi: 10.1145/3600211.3604712.

[15] P. Esmaeilzadeh, T. Mirzaei, and S. Dharanikota, "Patients' Perceptions Toward Human–Artificial Intelligence Interaction in Health Care: Experimental Study," *J. Med. Internet Res.*, vol. 23, no. 11, p. e25856, Nov. 2021, doi: 10.2196/25856.

[16] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento, "BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions," *Artif. Intell. Med.*, vol. 127, p. 102285, May 2022, doi: 10.1016/j.artmed.2022.102285.

[17] P. Pataranutaporn, R. Liu, E. Finn, and P. Maes, "Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness," *Nat. Mach. Intell.*, vol. 5, no. 10, pp. 1076–1086, Oct. 2023, doi: 10.1038/s42256-023-00720-7.

[18] J. L. Feuston and J. R. Brubaker, "Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–25, Oct. 2021, doi: 10.1145/3479856.

[19] E. Bondi *et al.*, "Role of Human-AI Interaction in Selective Prediction," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 5, pp. 5286–5294, Jun. 2022, doi: 10.1609/aaai.v36i5.20465.