Structured Medical Dataset Analysis Tool Based on ChatGPT

JinCheol Park
Medical AI Researcher Team
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
parkrealsteel@gmail.com

Jahyun Nam
Mediv
Chungcheongbuk-do, Rep. of Korea
namjahyunneo@gmail.com

Jeewoo Choi
Mediv
Chungcheongbuk-do, Rep. of Korea
chjoo91@gmail.com

Yong-Goo Shin
Department of Electric and Information Engineering, Korea University
Sejong-si, Rep. of Korea
ygshin92@korea.ac.kr

Seung Park
Department of Biomedical Engineering
Chungbuk National University Hospital
Chungcheongbuk-do, Rep. of Korea
Spark.cbnub@gmail.com

***Abstract*— This study proposes a medical data analysis tool using Chat Generative Pre-trained Transformer (ChatGPT), a natural language processing model (NLP), to directly communicate with medical data for various analyses. The tool recommends optimal AI models in terms of accuracy and efficiency using Shapley Additive explanation (SHAP) and showed average accuracies of 90.6% and 82% in the data analysis and AI model analysis stages, respectively. The proposed method showed the potential of chatGPT to revolutionize medical data analysis.**

Keywords—chatGPT, medical data analysis, natural language processing, deep learning

I. INTRODUCTION

As the importance of medical data analysis continues to grow, the need for processing large datasets becomes more prominent [1]. However, medical professionals often face difficulties in utilizing data analysis tools due to their complexity and technical challenges [2]. Previous methods for medical data analysis relied on set data analysis, AI model execution, and fixed output prediction, limiting their flexibility [3]. To address these issues, this paper proposes a new tool that utilizes the recently popular interactive AI model, chat Generative Pre-trained Transformer (chatGPT) [4]. Although the proposed method was trained for general purposes, not medical informatics, the flexibility and ability to easily search for desired information of the proposed method make it an ideal tool for medical data analysis.

By using the proposed method, users can communicate and directly interact with medical data, utilizing it for greater flexibility and a wider range of analyses. This study involves a total of four stages, using the proposed method seven times: four times for the data analysis stage, and three times for AI model analysis stage. In the data analysis stage, the proposed method was used four times: (1) analyzing the entire data uploaded through Streamlit [5], (2) providing explanations to users through natural language processing of the analyzed data, and (3) analyzing additional uploaded columns, and (4) recommending 7 AI models for data classification and analysis.

In AI model analysis stage, the proposed method was used three times: (1) providing explanations on AI analysis results, (2) processing Q&A about AI models through chat implemented in Streamlit, and (3) introducing survival status and survival probability, and explaining why such calculations were made when data were fed into an AI model trained by the proposed method. The proposed analysis tool provided appropriate analysis methods, optimal AI models, and data explanations for users without medical expertise. The optimality of the proposed AI model was determined using metrics such as accuracy (ACC), F1 score, Receiver Operating Characteristic Curve (ROC curve), and Shapley Additive explanation (SHAP) [6].

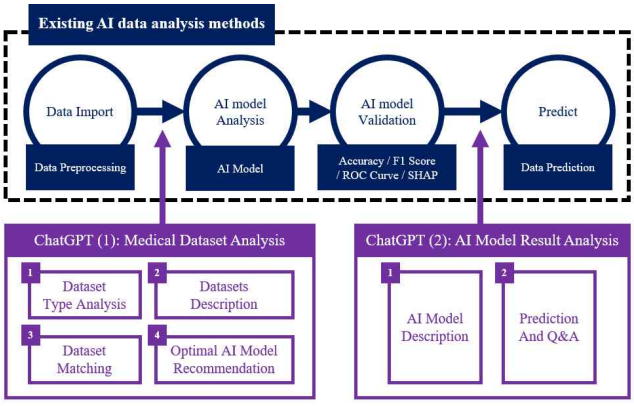


Fig. 1: The proposed methodology incorporates ChatGPT into the existing medical data analysis and AI model evaluation pipeline, as shown in the illustration. The integration of ChatGPT enables the analysis of diverse medical data and optimizes the AI model, which are challenging to achieve with conventional AI analysis methods alone.

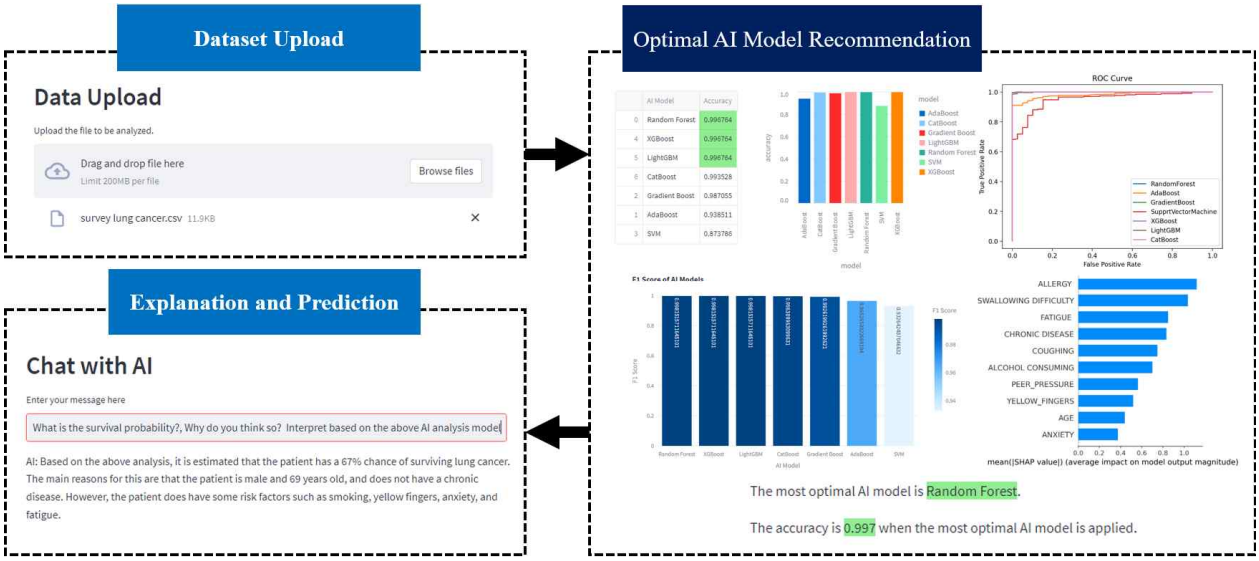


Fig. 2: The proposed method presents easy-to-use data analysis tools tailored for non-experts in the medical field. By simply uploading a medical dataset, the proposed method automatically conducts optimal analysis, enabling users to easily obtain predictions and explanations through a chat interface.

This study aims to prove the potential of the proposed method in medical data analysis and increase accessibility for medical professionals.

II. Comparison with Previous Methods

The complexity and significance of medical data analysis have driven extensive research toward automating this process [7][8][9][10]. This paper introduces a medical data analysis tool that automates a series of processes, including processing various medical data, identifying data files, and facilitating AI analysis. This methodology shares similarities with existing techniques, which motivates the comparative evaluation of its efficacy against two other methods in Section II.

The two previous methods chosen for comparison are similar to the proposed tool in that they are medical data analysis tools that automate processes for AI analysis by processing various medical data and identifying data files. The efficacy of the proposed and existing methods is compared and evaluated, including the existing method A that is specialized in data pre-processing and uses AutoML [11] technology to perform pre-processing for AI utilization. However, since it only performs pre-processing, manual analysis by the user is required in the subsequent steps (see Table 1). Conversely, existing method B applies NLP technology to automatically analyze medical data and perform AI analysis, yet has two problems. Firstly, the NLP analysis technique used to analyze randomly uploaded data has limitations in terms of time and cost, resulting in performance degradation. Secondly, AI analysis is limited to pre-established support vector machine (SVM) models, and tuning such as parameter modification is not possible.

In contrast, the proposed method using ChatGPT can address many of the aforementioned issues. Chat GPT is an NLP-optimized model that enables easy and fast data analysis and can flexibly handle undefined medical data. Moreover, the analysis tool proposed in this paper avoids using pre-set AI models and parameters, instead utilizing models recommended by ChatGPT for analysis, enabling higher accuracy and more flexible analysis than existing methods. Additionally, since ChatGPT participates in and learns a series of processes from medical data upload to prediction, the process can be fully automated.

TABLE I. Comparison between Previous Method and The Proposed Method

Dataset	Data Preprocessing	Analysis of Uploaded Data files	Description of Uploaded	Data Matching	AI Analysis
Previous Method A [7][8][10]	O	X	X	X	X
	· It automatically conducts data preprocessing.				
Previous Method B [9]	X	O (Used NLP)	X	X	△ [a]
	· Unstructured data analysis via NLP is costly and time-consuming. A solution is a semi-automated analysis tool that involves human participation. · The above approach allows basic analysis of uploaded datasets, but not other analytical techniques.				
The Proposed Method	△ [b]	O (Used GPT)	O	O	O [c]
	· Fast and flexible NLP analysis · Fully automated with Chat GPT				

[a.] Analysis was performed using the specified AI model, and the study used a Support Vector Machine (SVM) model.

[b] System solely executes rudimentary data preprocessing tasks, encompassing uppercase and lowercase transformation, numerical processing, and elimination of special characters, outliers, and missing values

[c] Rather than using a designated AI model, Chat GPT is used to read data and select and analyze the optimal model with the highest accuracy among 7 recommended AI models.

III. DATASETS FOR MEDICAL DATA ANALYSIS

In this study, a total of two datasets were used. The first dataset was publicly available data regarding the survival [12] of patients diagnosed with lung cancer, obtained from the AI learning platform Kaggle. This dataset consists of 309 instances and includes 16 features such as age, gender, allergies, chest pain, and lung cancer survival. In addition, modifications such as replacing certain words with synonyms, changing the order of feature values, or deleting target values were made for data analysis using the proposed method and for data matching experiments. Furthermore, to conduct matching experiments with other types of data, different datasets were used. The second dataset was publicly available data regarding the survival [13] of patients diagnosed with breast cancer, also obtained from Kaggle. Two datasets for cancer survival from Kaggle have 16 features. The second dataset for lung cancer survival consists of 4024 instances, but only the top 309 instances were imported for effective comparative analysis. It also includes 16 features such as age, survival status, and information about cancer cells. These two datasets are similar, and experiments were conducted to investigate whether matching between the two datasets is possible.

IV. METHOD

A. How to use the proposed method function

As shown in Fig. 3, there are two methods to leverage the capabilities of the proposed method.

The first method is to have the proposed method interact with users through natural language processing (NLP) analysis. This approach receives user requests in the form of chats, analyzes them through NLP, and uses the resulting outputs to build AI models. The analysis tool allows users to adjust the parameters according to their specific requirements to achieve the desired results.

The second way is to have the proposed method interact with the AI model through code analysis. This approach facilitates more flexible analysis without sticking to rigid frameworks or producing only fixed outputs. As part of obtaining the results, the proposed method is called to analyze the progress or results, and the results are then sent to the user. The proposed method was utilized for analysis of the features of the uploaded data files or to determine data matches.

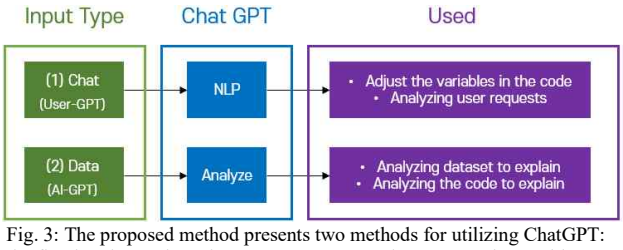


Fig. 3: The proposed method presents two methods for utilizing ChatGPT: the first involves using ChatGPT as an NLP tool to communicate with users, while the second involves using it as an analyzer to communicate with an AI model.

B. Dataset upload and dataset analysis

The process of uploading and analyzing a dataset can be separated into three steps. The first step is to implement the upload button using Streamlit, a web-based application, and preprocess the data for subsequent analysis in the proposed method. At this stage, standard data preprocessing techniques such as handling null values are applied. To optimize computational cost and speed, as well as increase accuracy by providing key information, we selected only the first five rows of data.

In the second step, we compare the new dataset to the previous one to make sure the columns and data match. To group similar columns during cross-column analysis, we preprocess the text and convert it into vectors. We use cosine similarity to calculate the similarity between vectors and group the columns based on a predefined similarity score [14]. If the proposed method finds it feasible, it merges similar columns to create a new analysis file.

The third step is to use the proposed method to analyze the data type of the uploaded dataset, consider supervised and unsupervised learning approaches, and recommend suitable AI models.

C. AI model training

After uploading the dataset, the data is analyzed by applying the AI model presented in the previous step. Various performance metric data, including accuracy, F1 score, ROC curve, and SHAP, are collected as a means to focus analysis and identify better AI models. An example of the performance results can be demonstrated in Fig. 2.

D. AI model analysis and prediction

The chat functionality was implemented using Streamlit, which combines natural language generation and understanding capabilities of the proposed method to enable three unique features:

- Natural language output of artificial intelligence model results: This function enhances user accessibility by outputting AI model results in natural language. This feature takes advantage of natural language generation capabilities of the proposed method to generate consistent and understandable sentences.
- Q&A conversation about AI model results: Users can ask questions related to AI model results and receive corresponding answers generated by the proposed method. This feature leverages natural language understanding capabilities of the proposed method to parse and understands user queries.
- AI Model Predictions via Data Upload and Analysis: This feature allows users to upload data via the proposed method and then use a pre-trained AI model to generate predictions. This feature relies on the ability of the underlying AI model to analyze the data and provide accurate predictions.

V. EXPERIMENTATION & DISCUSSION

A. Performance verification method

The article describes a performance verification method for ChatGPT, which is important to ensure the model produces the intended results. The proposed method is analyzed seven times, with a minimum of 100 verifications per item. The verification process includes data analysis, description provision, data matching, AI recommendation, explanation of AI analysis results, Q&A, and data prediction. The article aims to validate the accuracy of the proposed method in both data analysis and AI model analysis stages.

B. Data analysis & Description provision

Data analysis involves identifying the data columns, determining the target variable, and characterizing the data. Thus, it is imperative to perform feature analysis and provide explanations to gauge the quality of the analysis and interpretation. Subsequently, we assessed the adequacy of feature analysis and analysis output by verifying their proper implementation.

TABLE II. DATA ANALYSIS & DESCRIPTION ACCURACY

Dataset	Feature Analysis Accuracy [f]	Analysis Output Accuracy [g]
Supervised Learning Dataset (1) [d]	1	0.94
Unsupervised Learning Dataset (2) [e]	1	0.82

[d.] This is a survival dataset for lung cancer patients. It includes 16 columns, and the target feature is the survivability of the patients.

[e.] This is a modified dataset of the supervised learning dataset from (1), where the target feature has been removed.

[f.] The proposed method demonstrated a 100% recognition rate for extracted features from uploaded data, irrespective of whether the data was supervised data or unsupervised data for both feature learning paradigms. It also did a good job of explaining the recognized features.

[g.] Regarding the request to elucidate the characteristics of the uploaded data based on feature recognition by the proposed method, there was a discrepancy in the success rates between supervised and unsupervised learning data.

Table II. (1) shows that feature and target analysis yielded satisfactory results for all 50 datasets, irrespective of the data type, as demonstrated in (1). Nevertheless, as depicted in (2) of Table II. (2), the proposed method sometimes provides brief responses without providing an in-depth explanation of the intended meaning during data analysis. This is particularly noticeable when analyzing datasets without any identifiable target values. It was found that recognizing the target in the process of analyzing the data of the proposed method affects the accuracy.

C. Data matching

Matching was verified according to the type of dataset by analyzing the matching between lung cancer survival data and various types of datasets. The success criterion for data matching was based on accurately matching the features of two data sets while disregarding any typos or abbreviations. Table III shows the data matching results.

TABLE III. DATASET MATCHING

Dataset [h]	Independent Variable		Matching Accuracy
(1) Same Data	Change feature		Possible: 1
(2) Same Data	Modify features with similar name		Possible: 1
(3) Same Data	Change feature + Modify feature with similar name		Possible: 0.92
(4) Same Data	More than 50% of different features		Error: 0.58
(5) Other Data	Another target	Supervised learning data	Impossible: 1
(6) Other Data	Another target	Unsupervised learning data	Impossible: 1

[h.] The dataset consists of survival data for lung cancer patients, with 16 columns and the target feature being patient viability. Matching experiments were conducted between this dataset and datasets (1) to (6).

In the case of (1) and (2), the matching of the dataset was well connected by deriving that the data were the same in the data analysis step. However, in the case of (3) where multiple changes were made to the same data, 4 out of 50 cases were recognized as different data. In the case of (4), when compared to the data in which manipulation was applied to more than 8 columns excluding the target, errors were output with a probability of 52% because it was difficult to determine whether it was the same data or different data. As in the case of (5) and (6), if the data differed only in the target variable, it was recognized as other data, and "Impossible" was output in response to the question of whether matching was possible.

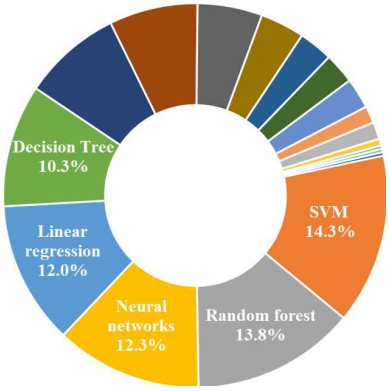
D. AI recommendation

The purpose of the data analysis phase is to analyze the data to identify the subject, determine the type of data, and recognize and interpret the uploaded data. Based on this interpretation, the proposed approach proposes seven AI models that are considered optimal for AI analysis using the analyzed data. The recommended AI model is different for each data set being analyzed, and the nature of GPT suggests different AI models for each run. Table IV shows the AI model recommended in the proposed method for the lung cancer patient survival dataset, and the actual accuracy of the proposed AI model was obtained through coefficients.

Support Vector Machine (SVM) was recommended in all 50 cases and was the most frequently recommended AI model. However, it had comparatively lower accuracy at 87.3% compared to other models. On the other hand, although Categorical Boosting (CatBoost) received only five recommendations, it achieved a remarkably high accuracy of 99.3%, which is comparable to the highest accuracy achieved by Random Forest at 99.7%. Therefore, it was confirmed that the proposed method recommends the most general model rather than the most accurate model for data analysis. To output the optimal AI model, it was analyzed that it is most desirable to output the AI model with the highest ACC among the models recommended in the proposed method. For this reason, the proposed data analysis tool trains the AI model with the highest ACC among the 7 recommended models to output the optimal AI model. The process of finding the model with the highest ACC is shown in Fig 2

TABLE IV. AI RECOMMENDATION

Graph recorded cumulatively [i]



Recommended AI model [j]	Count	Recommend Score	AI Model Accuracy
Support Vector Machine (SVM) [k]	50	1	0.873
Random Forest	48	0.96	0.997
Linear Regression	42	0.84	0.400
Decision Trees	36	0.72	0.997
Logistic Regression	26	0.52	0.903
Gradient Boosting	19	0.38	0.987
Naïve Bayes	13	0.26	0.906
Extreme Gradient Boost (XGBoost)	10	0.20	0.997
K-Nearest Neighbor	9	0.18	0.906
Light Gradient Boosting Machine (LightGBM)	5	0.10	0.997
Categorical Boosting (CatBoost)	5	0.10	0.993
Adaptive Boosting (AdaBoost)	1	0.02	0.939

[i] *The method proposed in this study recommends seven optimal artificial AI models by utilizing the analysis of medical data in the previous step. The selection of recommended AI models is data-specific, and the nature of the proposed method architecture suggests different AI models for each run.*

[j] AI models that did not display ACC due to configuration errors were omitted from the table.

[k] The support vector machine (SVM) was the most frequently recommended AI model, being recommended in 50 out of 50 cases in the AI model recommendation process. However, its actual accuracy was found to be low. Therefore, this paper proposes an improvement by selecting the AI model with the highest accuracy among the seven recommended models.

E. *Performance verification of AI model description proposed after AI model analysis.*

The interaction with the proposed method for the AI model was verified. As an index to evaluate whether the correct explanation was derived, three essential keywords that must be included in each answer were selected, and the evaluation was based on whether or not the three essential keywords were correctly derived.

TABLE V. EXPLANATION OF AI ANALYSIS RESULTS

Chat Input	Whether to output requested key keywords			Summing Accuracy
AI model Result Output	AI Model Name Accuracy	AI Model Description Accuracy	AI Model Accuracy	0.71
	1	0.12 [l]	1	

[l] The explanation of the specific AI model used for analysis was often unsatisfactory, as it frequently provided a general overview of machine learning and deep learning, rather than a detailed explanation of the selected AI model.

In Table V, the verification of whether the AI model was explained well was based on the AI model name, AI model description, and accuracy output when applying the AI model. Since there was a process of recommending and utilizing an AI model in the data analysis stage, the name of the AI model and the proposed method for ACC show that the output was successful. However, there were many cases in which only general explanations about machine learning and deep learning were repeatedly output when requesting an explanation of the AI model itself, rather than giving specific explanations about specific AI models. Therefore, it is deemed more efficient to gather information about AI models using crawling than by listening to explanations through the proposed method.

F. *Performance verification of AI model description proposed after AI model analysis.*

To evaluate the accuracy of the predictions, we selected three essential keywords that must be included in each answer and assessed whether the correct explanation was derived based on the presence of these keywords. We conducted two experiments: in the first, we input data for all columns in the uploaded file, while in the second, we used only the 7 main features obtained through SHAP. Table VI shows the results of a random test dataset based on three criteria: correct prediction of lung cancer patient survival, output of survival probability, and output of the basis for the judgment. The prediction accuracy for survival was 96% when values were input for all columns, but lower accuracy was obtained when using only the 7 main features. This is likely due to a limitation of the dataset rather than an error in the AI model. The output of the survival probability was consistent across multiple experiments, indicating stable performance. Judgment Reason Output was mainly output based on the feature with high weight.

TABLE VI. DATA PREDICTION RESULTS

Input data to predict	Whether to output requested key keywords			Summing Accuracy
	Survival Prediction Accuracy	*Survival Probability Prediction Accuracy*	*Judgment Reason Accuracy*	
Input all column features	0.96	1	0.86	0.94
Input part of column features	0.56	1	0.88	0.81

VI. CONCLUSION

The proposed method in this paper enhances the inflexible data analysis method by enabling ChatGPT to engage in free-form conversations with users and allowing for interaction with AI models themselves. The method is convenient and easy to use, with analysis, explanation, and optimal settings generated automatically by uploading data and entering information. The proposed method exhibited an average accuracy of 90.6% in

the data analysis stage and 82% in the AI model analysis stage, demonstrating satisfactory performance. This method has great potential as a data analysis tool, with its unique conversational feature expected to expand the horizon of data analysis. Plans include extending analysis to both structured and unstructured data.

ACKNOWLEDGEMENTS

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1074070021).

REFERENCES

[1] Wan, Thomas TH. "Healthcare informatics research: from data to evidence-based management." Journal of Medical Systems 30 (2006): 3-7.

[2] Nam, Jahyun, et al. "A BERT-Based Artificial Intelligence to Analyze Free-Text Clinical Notes for Binary Classification in Papillary Thyroid Carcinoma Recurrence." 2023 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2023.

[3] Gheisari, Mehdi, Guojun Wang, and Md Zakirul Alam Bhuiyan. "A survey on deep learning in big data." 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC). Vol. 2. IEEE, 2017.

[4] Streamlit Developers. Streamlit Documentation. (https://docs.streamlit.io/en/stable/), accessed September 10th, 2021.

[5] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

[6] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017)

[7] Waring, Jonathan, Charlotta Lindvall, and Renato Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare." Artificial intelligence in medicine 104 (2020): 101822.

[8] Mustafa, Akram, and Mostafa Rahimi Azghadi. "Automated machine learning for healthcare and clinical notes analysis." Computers 10.2 (2021): 24.

[9] Marshall, Iain J., and Byron C. Wallace. "Toward systematic review automation: a practical guide to using machine learning tools in research synthesis." Systematic reviews 8 (2019): 1-10.

[10] O'Byrne, Ciara, et al. "Automated deep learning in ophthalmology: AI that can build AI." Current Opinion in Ophthalmology 32.5 (2021): 406-412.

[11] He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art." Knowledge-Based Systems 212 (2021): 106622.

[12] Hong, Z.Q. and Yang, J.Y. "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane", Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991.

[13] JING TENG, January 18, 2019, "SEER Breast Cancer Data", IEEE Dataport, doi: https://dx.doi.org/10.21227/a9qy-ph35. https://ieee-dataport.org/open-access/seer-breast-cancer-data

[14] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)