Journal of Computer Technology and Software
ISSN:2998-2383
Vol. 2, No. 3, 2023

Improving Medical Diagnosis with Artificial Intelligence: Naive Bayesian Classification Algorithms

Lysander Cole

Department of Computer Science, University of Maryland, USA

lysander.cole39@umd.edu

Abstract: Artificial intelligence (AI) has the potential to revolutionize medical diagnosis by imitating and expanding human intelligence. This paper explores the integration of AI into hospital information systems to address key challenges in the medical field, such as insufficient medical resources, lengthy doctor training cycles, high medical costs, and misdiagnosis rates. We propose an intelligent diagnostic system leveraging Naive Bayesian Classification and symptom screening algorithms to improve accuracy and efficiency. The system dynamically updates disease diagnoses based on symptom inquiries and provides effective symptom information for final diagnoses. The continuous development of AI technologies, including computer vision, natural language processing, and speech recognition, further enhances the capabilities of medical AI systems. This integration promises to optimize resource use, elevate treatment levels, and transform the traditional medical industry, paving the way for comprehensive applications in clinical research and patient care.

Keywords: Dialogue System; Symptom Screening; Disease Diagnosis; Naive Bayesian Classification.

1. Introduction

Artificial intelligence, as a hot discipline, mainly studies and develops a new technology for imitating and expanding human intelligence, including principles of things, ways of thinking, and special skills [1]. The goal of introducing AI technology is to reduce a large amount of human resources and runtime. In addition, the deep reinforcement learning of AI technology can dealing with large-scale data and provide results in a short period of time.

Intelligent analysis has real- time data processing and prediction capabilities [2]. In recent years, the quality and efficiency of medical services have always been the focus of people's attention.

In response to the four major problems faced by the international medical field: insufficient medical resources, long doctor training cycles, high medical costs, and high misdiagnosis rates of doctors[3], the combination of artificial intelligence technology and hospital information systems can not only effectively solve the above four problems, but also strengthen hospital information systems, The modern hospital information technology is gradually transitioning from traditional hospital information systems to intelligent hospital information systems. Background AI intelligent diagnostic system is kind of the expert system in the medical field, which is an important branch of artificial intelligence technology applied in the field of medical diagnosis[4].utilizing AI technology to transfer the knowledge of medical experts to the system and using the system to simulate the professional knowledge and reasoning judgments of medical experts, Intelligent diagnostic system like doctors, professional diagnosis can be made based on the clinical symptoms of the patient's disease and effective treatment and prevention plans can be proposed.

The theoretical foundation of intelligent diagnostic system involves all disciplines in the medical field. Therefore, a medical knowledge base that can store pathological and physiological mechanism description models and medical knowledge of expert doctors is the main foundation of intelligent diagnostic system.

The intelligent diagnostic system is composed of a knowledge base, inference, and algorithms. There are a large number of unstructured or semi-structured factors between medical disease diagnosis and clinical symptom manifestations. The use of conventional diagnosis systems requires the professional knowledge of doctors to make manual judgments to determine patient diagnosis and treatment plans, while the intelligent diagnostic system only requires logical inference through algorithms, By fully utilizing medical knowledge base and inference technology, misdiagnosis caused by doctors' experience issues can be avoided and the accuracy of diagnosis can be improved.

2. Architecture

Generally, the architecture of a modern intelligent diagnostic system include five basic modules[5]. 1) Medical Knowledge Base: used to store and manage pathological and physiological institution description models and medical knowledge of expert doctors, which includes authoritative medical textbook knowledge, common sense medical knowledge, and empirical knowledge of medical experts. The quantity and quality of medical knowledge in the knowledge base are the core factors for evaluating the performance and problem-solving ability of intelligent consultation systems. 2)Database: used for store raw data, factual data, and inference results in the medical field. 3)Knowledge Acquisition Module: It is a medical knowledge base used for obtain, modify, expand, improve, and maintain medical knowledge from medical experts. 4)Inference Machine Module: based on the treatment needs of patients' clinical symptoms and the raw data in the medical knowledge base, algorithms are utilized to match the symptoms of the medical knowledge base and perform inference to confirm the diagnosis of patients' clinical symptoms. As the knowledge in the medical knowledge base often cannot fully cover all clinical symptoms, it is necessary to coordinate and control the operation of the entire system for preliminary inference. 5)Explanation and input output interface: It is a human-computer interaction interface that provides explanations of the reasoning basis and route when necessary, providing convenience for doctors to understand the reasoning process and maintain knowledge bases and other systems. Figure 1 is a schematic diagram of an artificial intelligence medical diagnosis system.

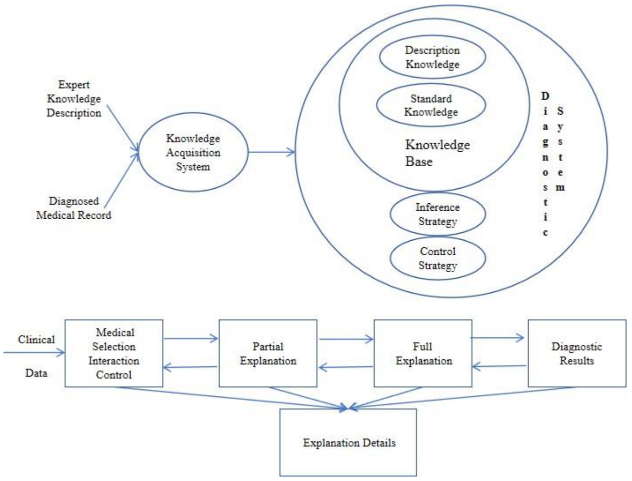

Fig 1. Schematic diagram of an artificial intelligence medical diagnosis system

As shown in the figure, the system is composed of a knowledge base and inference control separately, which in turn consists of disease Description Knowledge and Standard Knowledge. Description Knowledge describes the disease process and patient's disease data structure as a structured way to store relevant information. Standard Knowledge refers to the decision inference rules or logical constraints used for semantic network nodes or framework structures, thus, data can be effectively inferred and interpreted.

Inference and control strategies are referred to as "inference machines", which generate and manage possible hypotheses during the inference process, used to infer and interpret information. By comparing input data with predefined patterns or rules, potential associations and rules are discovered. The system continuously scans and updates input data to capture new information and changes.

In the above figure, after preliminary analysis of the patient's symptoms, the system can provide some preliminary explanations about possible diseases or diagnoses. Then, the system compares these preliminary explanations with the knowledge base we have to obtain more accurate explanations, summarize conclusions about the patient's disease situation, and provide relevant treatment suggestions.

3. Basic Technologies and Methods

As an important scenario in smart healthcare, the technology of Artificial intelligence has developed quickly. Currently, the core technology of the latest AI intelligent diagnostic system include Medical Dialogue Diagnosis Technology, Medical Knowledge Graph Technology, "Inference Machine" Technology, Disease Diagnosis Algorithm, Symptom Screening Algorithm, and Naive Bayesian Classification. Continuously improve the intelligent consultation system based on the above technologies combined with AI algorithms.

3.1. Medical Dialogue Diagnosis Technology

Medical Dialogue Diagnosis Technology is the application of task-based dialogue technology in the medical field. It is a medical dialogue system based on natural language processing, computer vision, and speech technology. The subsystem defines replies in advance in the system, and the dialogue system uses methods such as rule engines, pattern matching, or machine learning classifiers to filter out the best reply from the knowledge base and output it to the user during the dialogue, It is equivalent to a knowledge search task. The implementation of a retrieval based dialogue system requires the preparation of a rich knowledge base. After receiving user input, the system retrieves and selects the answer content from the knowledge base. Due to the predefined nature of responses, the advantage of a retrieval based question answering system is its high quality of answers and smooth and natural expression. However, due to the fact that its answer content is entirely based on the knowledge base, its disadvantage is that it has high requirements for the richness of the knowledge base. In addition, the rules cannot be exhaustive and are not easy to expand, and do not have intelligence. Compared to retrieval based dialogue systems, generative dialogue systems do not have predefined responses, but are trained based on a large amount of high-quality topics. Their advantage is that they allow users to input arbitrary sentence structures and cover a wide range of topics; The disadvantage is that the quality of the generated reply sentences cannot be guaranteed, and there may be fluency issues or low-level grammar errors such as syntax errors. The main methods of medical dialogue diagnosis technology include pipeline method and end-to-end method. In AI intelligent diagnostic system, pipeline method is mainly used, which includes the following five functions:

1) Speech Recognition Module: responsible for recognizing user input signals as corresponding text.

2) Natural Language Understanding Module: responsible for converting the text information recognized by speech recognition into a format that the computer can understand for representation.

3) Dialogue Management Module: responsible for the dialogue management function, which includes two parts: dialogue state tracking (DST) and dialogue strategy (DP). It judges the current dialogue state based on the output of the natural language understanding module, and makes decisions on the execution actions that should be taken in the current dialogue state.

4) Natural Language Generation Module: responsible for

converting the execution actions output by the dialogue management module into natural language text for output.

5) Speech Synthesis Module: responsible for converting the output natural language text into speech form for output.

3.2. Medical Knowledge Graph Technology

The knowledge graph is a combination of specialized knowledge and thinking patterns of doctors, which includes knowledge content, knowledge volume, and professional logical relationships between knowledge. The logical application of these knowledge can interpret and judge disease information to make correct judgments and decisions. Therefore, the medical knowledge graph, the most important thing is not simply a list of superficial connections, but a structured medical knowledge point (information, data) and its correlation that conforms to the inherent logical mechanism of medical knowledge. Based on the medical knowledge, the relevant algorithms be clearly and reliably explained, and meet the objective requirements of clinical medicine that must be verified. The impact of knowledge graph on algorithms are:

1) The knowledge graph defines how to structure the disease information and data that needs to be processed. Only structured information and data can meet the requirements of large-scale computing.

2) The knowledge graph defines the logical relationships between disease information and data, determining the flow and order of disease information/data in artificial intelligence operations. Programs use these flow and order to simulate doctors' cognition and reasoning of information and data, determining the level of intelligence in artificial intelligence.

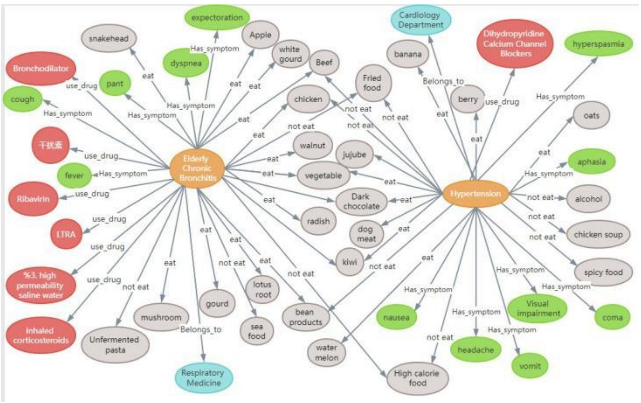

Fig 2. The knowledge map of chronic bronchitis and hypertension in the elderly

The conventional data types of medical knowledge graph include label nodes, disease nodes, and relationship types. The tag nodes include: Disease, Department, Drug, Food, Symptom, etc. The Disease node contains the name, introduction, prevention, treatment, and treatment period. Relationship type which is the connection between nodes: belongs_ To, has_ Symptom, use_ Drug, do_ Eat (suitable for eating), can_ Not_ Eating and other information. For example, the knowledge map of chronic bronchitis and hypertension in the elderly.

3.3. "Inference Machine" Technology

The essence of "Inference Machine" Technology is a deep reinforcement learning technology that combines deep learning and reinforcement learning methods, it is aimed to train intelligent agents to make decisions and actions in complex environments. It is based on the framework of reinforcement learning and uses deep neural networks as approximation of functions to handle high-dimensional input and output spaces.

In deep reinforcement learning, agents could learn through interaction with the environment and optimize their behavioral strategies through continuous trial and error. It influences the environment by observing its state, selecting actions, and updating its strategies based on feedback (rewards or punishments) from the environment. Deep neural networks could execute approximate value functions or policy functions to assist intelligent agents make more accurate decisions.

3.4. Symptom Screening Algorithm

Symptom Screening Algorithm is developed based on the differences in symptom sets. symptom screening is the core of conversational medical diagnosis. In the field of medical dialogue, medical conversation agents are important tools that help doctors respond to common diseases that are easy to diagnose and collect medical information from patients with difficult diseases [6]. The algorithm introduces the idea of excluding diseases with the highest probability of illness in the real medical diagnosis process, combined with the idea of comprehensive search, and through the intersection and union operation of the disease symptom set, it can filter the symptoms that could distinguish diseases better with fewer inquiry rounds, providing a decision-making basis for the final Naive Bayesian Classification.

The algorithm utilizes the conditions of known positive and negative symptom sets to select symptoms with high discrimination ability to distinguish different diseases in the disease set D as the core idea. Implement the following functions in the symptom screening algorithm:

1) Based on known symptom information, infer the possibility of each disease in the list of confirmed diseases.

2) Ask about the symptoms with a higher probability of occurrence among other diseases with a higher probability of illness

3) Minimize the number of symptoms that need to be inquired about as much as possible.

3.5. Disease Diagnosis Algorithm

The real hospital's disease diagnosis process can be divided into three steps:

(1) the patient presents their own discovered symptoms.

(2) Doctors communicate with patients to confirm the existence of other symptoms.

(3) The doctor determines the final type of disease based on the initial symptoms raised by the patient and the symptom information obtained through communication

(4) Wherein step (2) is the symptom screening process, it is aimed to select the most appropriate symptom to minimize the number of rounds of dialogue and provide more basis for the final judgment. Step (3) can be considered as a classification task, the disease diagnosis algorithm utilizes the differences between symptom's data sets of various diseases to select symptom options that can relatively distinguish these diseases in batches and rounds. It has the characteristic of high screening efficiency. Finally, it is necessary to confirmed updating the list of diseases dynamically in order to determine whether the known symptom information is sufficient to support diagnostic decisions, this is to support whether the intelligence diagnosis system could provide final diagnose directly or

continue running the symptom screening algorithm to obtain additional symptom information. This organically combines disease diagnosis algorithm and symptom screening algorithm, achieving the entire process of intelligent diagnosis.

Algorithm 1: Symptom Screening Algorithm

Input: The positive symptom set is S^+, the negative symptom set is S^-, and the disease set to be confirmed is D_0
Output: Symptom S

1. Calculate the probability of each disease D_0 under the conditions of known positive symptom set S^+ and negative symptom set S^-, and sort it in descending order of probability, with a length of $D_0 = m_0$.
2. Take the union of the symptom sets of the top m diseases in set D_0, and define this union as set S_U, with D_a as the set of m diseases merged.

$$S_\cup = \bigcup_{d_i \in D_0} D_i$$

The constraint condition is:

(1) The union S_U of symptoms cannot be a complete set.
(2) The number of m can only be less than or equal to the half of the number of diseases m_0 in the set $D_{0.}$
(3) Under the condition of b, choose the maximum value of m as much as possible.

$$S_\cup \neq S, m \leq \frac{m_0}{2}, m = max(m)$$

3. Based on the m diseases selected in the second step, construct a set D_a; Remove D_a from set D_0 to obtain set D_b, where D_b=D_0- D_a. For the set D_b, find the maximum intersection of the symptom sets of n diseases, denoted as $S_\cap$。

$$S_\cap = \bigcap_{d_i \in D_b} D_i$$

The constraint condition is:

(1) The selected n diseases and their symptom intersection $S_\cap$ is not an empty set.
(2) Choose as many diseases as possible.

Therefore, when the symptom sets of each disease in the set D_b do not intersect with each other, we can choose any symptom set of the disease as $S_\cap$

$$S_\cap \neq \emptyset \ , n = \max(n)$$

4. The difference between the symptom sets $S_\cap$ and $S_\cup$ is marked as S.
5. Select the symptom S with the highest probability from S_0.

Algorithm 2: Disease Diagnosis Algorithm

Input: Self Diagnosis Description
Output: Diagnosis of Diseases

1. Extract positive and negative symptom sets, namely S^+ and S^-, from the patience's description.
2. Update the list of confirmed diseases D_0 based on S^+ and S^-. If there are multiple diseases in D_0, symptom screening is required to select one symptom S that has not been inquired about. If the length of D_0 is less than or equal to 1, the result will be directly classified and output.
3. Ask the patient if there are symptoms S selected by the screening algorithm。
4. Based on the patient's response, update S+and S -, and then repeat step 2.
5. Based on S^+ and S^-, use Naive Bayesian Classification methods for classification and output the final diagnostic results.

Based on the above assumptions, we can improve the process of disease diagnosis algorithms, as shown in Figure 3.

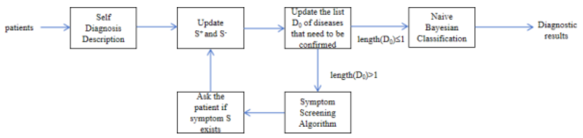
patients
Self Diagnosis Description
Update S⁺ and S⁻
Update the list D₀ of diseases that need to be confirmed
length(D₀)≤1
Naive Bayesian Classification
Diagnostic results
length(D₀)>1
Symptom Screening Algorithm
Ask the patient if symptom S exists

Fig 3. The process of Disease Diagnosis Algorithm

The process of this algorithm is divided into the following steps:

1. Based on the patience's self-description, extract the positive symptom set $S^+$and negative symptom set S^-, and store the existing symptoms of the patience in S^+, and the non-existent symptoms in S^-.
2. Update the list of confirmed diseases D_0 based on $S^+$and S^-. If only one disease remains, skip to step 5. Otherwise, conduct symptom screening, and select the symptom S to be inquired.
3. Ask the patience if there is a symptom S.
4. Update the positive symptom set $S^+$and negative symptom set S^- based on the patience's response. If symptom S exists, add it to S^+; if not, add it to S^-. Then return to the second step.
5. Input the positive symptom set $S^+$and negative symptom set S^- to the Naive Bayesian Classification module for classification, output the disease with the highest probability as the diagnostic result, and output the result.

The algorithm is a Disease Diagnosis Algorithm based on Naive Bayesian Classification[7]. This algorithm takes Naive Bayesian Classification as the core diagnostic link, integrates symptom inquiry process, organically combines symptom inquiry and disease diagnosis classification, and realizes the entire process of intelligent diagnosis.

4. Evaluation

According to the three steps of disease diagnosis and their functional objectives, the evaluation indicators for the effectiveness of disease diagnosis are diagnostic success rate and conversation rounds

(1) Diagnostic Success Rate.

The diagnostic success rate is equivalent to the prediction accuracy in the classification task, as shown in below:

$$success = \frac{num(success\ dialogue\ )}{num(all\ dialogue)}$$

(2) Conversation Rounds.

Due to the background of this task being a dialogue system, the efficiency of the dialogue determines the intuitive performance of the dialogue system, and the number of dialogue rounds is an important indicator of the performance of the dialogue system

$$turn\_rate = \frac{|new\_turn - raw\_turn|}{raw\_turn}$$

In this task, regardless of whether the doctor needs to inquire about other symptoms, there must be at least one complete conversation round, that is, at least one conversation round. In addition to the number of conversation rounds, in order to better quantify the performance of the model in the number of conversation rounds, the relative percentage of conversation rounds is to measure the relative change in the number of conversation rounds after using disease diagnosis algorithms, which is to calculate the overall number of conversation rounds and the number of conversation rounds in the dataset.

5. Clinical Application Effect

The AI intelligent diagnostic system is one of the most successful cases of utilizing AI technology into the medical sector, simulating the diagnosis and treatment of diseases by medical experts. After research on most hospitals using AI intelligent consultation systems, it was found that the accuracy of disease diagnosis and dialogue efficiency have significantly improved in real medical scenarios by utilizing medical dialogue diagnosis technology, medical knowledge graph technology, "inference machine" technology, disease diagnosis algorithms The symptom screening algorithm and naive Bayesian classification. Even if the number of symptom inquiries is small, the system can still provide more than 90% diagnostic accuracy, which is of great significance for clinical practice.

At present, all large hospitals around the world are gradually introducing intelligent diagnostic system to improve their medical and service levels, as the system can quickly and accurately determine the type of disease patients suffer from and provide corresponding treatment plans. The significant effect is to automate the analysis of medical records, combine patient medical record data, medical knowledge base, and the latest medical development trends and other relevant information, conduct disease risk assessment and disease prediction, and predict the type and severity of the patient's disease. In order to achieve efficient patient diversion, patients can receive the most accurate disease assessment, effective treatment plans, and preventive measures at home. Compared to traditional medical diagnosis methods, intelligent diagnostic system based on artificial intelligence have the following advantages:

1. High Efficiency

By the support of artificial intelligence technology, the intelligent diagnostic system runs faster, and the medical diagnosis time can be shortened to a few seconds, the efficiency of medical treatment could be improved greatly. Moreover, it could simultaneously deal with the medical record data of multiple patients, thereby improving the efficiency of hospital resource utilization, which is able to solve the problem of medical resource collapse, and greatly improving the efficiency and quality of medical services.

2. High Accuracy

Traditional medical diagnosis methods are easily influenced by subjective factors from doctors, while intelligent diagnostic system based on artificial intelligence obtain accurate diagnostic results through a large amount of data analysis and model training. At the same time, it can also correct loopholes in medical records and avoid misdiagnosis caused by data errors.

3. High Degree of Automation

The intelligent diagnostic system based on artificial intelligence is a fully automated process that could automatically retrieve medical record data, analyze medical record records, generate diagnostic reports, and provide corresponding treatment plans. This not only improves medical efficiency, but also reduces the workload of medical staff, making them more focused on patient care and treatment.

The intelligent diagnostic system based on AI technology is still in a rapid development stage, and in addition to the points mentioned above, there are also some challenges:

1. Insufficient Data

AI intelligent diagnostic system requires a large amount of data for learning and training. If the data is insufficient or of low quality, it will affect the accuracy and effectiveness of diagnosis.

2. Data privacy and security

Big data is the foundation of artificial intelligence medical diagnosis, but the collection and use of data involve the protection of personal privacy. If data leakage occurs, it may directly harm individuals' privacy rights, and comparative technical measures need to be taken to strengthen data protection.

3. Alternative Doctor

Although artificial intelligence can be very effective in detecting and analyzing a large number of cases, it still requires human doctors to participate in diagnosis in some details. Some cases require emotional therapy for patients, and artificial intelligence is currently not fully developed. Therefore, a better balance should be sought between artificial intelligence technology and human doctors.

4. Technical credibility and reliability

The application of artificial intelligence technology in the field of medical diagnosis is still in the development stage, and there may be evaluation errors. It is necessary to conduct a comprehensive evaluation of core technologies such as system architecture, functions, and algorithms.

Therefore, AI intelligent diagnostic system is an emerging medical technology that requires comprehensive consideration of its advantages and disadvantages while ensuring data privacy and security, combined with the experience and judgment of human doctors, the system functions, algorithms, and medical knowledge base should be continuously improved to enhance the accuracy and efficiency of diagnosis.

6. Conclusion

With the scarcity of medical resources, the demand for intelligent medical systems is becoming increasingly urgent. In order to meet the growing demand of intelligent diagnosis. Firstly, using Naive Bayesian Classification and a symptom screening algorithm based on differences in disease symptom sets to solve the problems of low accuracy, long running time, and poor interpretability in conversational disease diagnosis. Specifically

1) A disease diagnosis algorithm based on naive Bayesian classification was designed by combining symptom inquiry with disease diagnosis classification. This algorithm dynamically updates the list of confirmed diseases in real scenarios, improving the interpretability and performance of the diagnostic process, and achieving the entire process of intelligent diagnosis.

2) A symptom screening algorithm based on the differences in disease symptom sets has been introduced to provide effective symptom information for the final diagnosis process. This algorithm can improve the accuracy and efficiency of conversational disease diagnosis with fewer questioning steps. By the continuous development of artificial intelligence technology and the smart hospital architecture, it is significant to utilize artificial intelligence in hospital scenarios, especially in medical diagnosis situation. The technology could save medical resources, and improve the hospital's treatment level, Meanwhile, it is easily to promote the upgrading of service and treatment level for the traditional medical industry, and alleviate the shortage of medical resources. Computer vision recognition technology, natural language processing technology, and speech recognition technology, as the three major engines of deep learning in artificial intelligence, are widely used in the development of medical artificial intelligence products. At present, the research hotspots and fields of medical artificial intelligence are still very broad. With the continuous deepening of understanding of medical artificial intelligence and the continuous development of related technologies, the future of medical artificial intelligence will inevitably involve the collection and organization of clinical information, data extraction and analysis, diagnosis and treatment quality control, physical examination screening, treatment plan decision-making, efficacy evaluation, and prognosis prediction. It will gradually be applied to the full chain medical scene of clinical research, moving from early diagnosis to clinical treatment, providing doctors with more and better assistance.

References

[1] Nianlin S, Xueling H, Pengpeng W.(2019). Research on the Application of Artificial Intelligence in the Medical Field. Nursing of Integrated Traditional Chinese and Western Medicine, 5(11), pp.141-143.

[2] Wenjuan Z.(2013). Research on Intelligent Statistical Analysis System for Hospital Information. Technological Information, (26), pp.279-280.

[3] Hongkun Y, hao H, Qiang L, Ziye Y.(2018). Expanding the boundaries of medical artificial intelligence in Xinjiang. Artificial Intelligence, 4(11), pp.89-97.

[4] Zhaofang Z, Lin A.(2009). Research on the Construction and Application of a New Traditional Chinese Medicine Expert System. JOURNAL OF LIAONING UNIVERSITY OF TCM, 11(10), pp.9-10.

[5] Huikang Z, Zongcai Q, Jinghui Q, Jun L, Weizhong X, Jinhua L.(2002). Expert System and Its Application in Medical Diagnosis. J Fourth Mil Med Univ, 4(S1), pp.73-76.

[6] Liu W, Tang J, Cheng Y, Wenjie L, Yefeng Z, Xiaodan L.MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation[J]. 2022.DOI: 10. 1007/ 978-3-031-17120-8_35.

[7] Lianhai Y, Xiangwen L, Jing X.(2020). Research on Spam Filtering Algorithm Based on Improved Bayesian Principle. Computer & Digital Engineering, 3(48), pp.513-517.