International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 1, No. 3, 2026

Application of Automated Data Extraction Technologies in Ophthalmic Electronic Health Records

Yinghai Yu

School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China

Abstract: Now that the healthcare industry is increasingly using digital technology, there is a lot of useful unstructured data in the electronic health record of the eye, but it is easy to make mistakes when it takes time to process it manually. In order to solve this problem, this paper carefully examines the application of optical character recognition and natural language processing in ophthalmic electronic medical records, and also pays special attention to the large language models that have recently been very hot. We checked a lot of data through Google Scholar, and after careful screening, we selected 30 good quality articles to study. From the results, these technologies have their own advantages and disadvantages: optical character recognition processing standardized equipment reporting accuracy is very high, but it is not very good when it comes to handwriting. Natural language processing can find vision data and disease characteristics from the medical record text, but if the medical record format is not uniform, the effect will be reduced. The latest large-scale language models are really powerful, they can directly handle words and pictures, and change the entire workflow, but they also have their own problems, such as data security, high operating costs, and possible hallucinations. In general, these automated technologies have deepened the use of data on electronic medical records of ophthalmology. Later research can focus more on how to better combine different types of data, and to develop some language models that are specific to ophthalmology, smaller, and can operate safely locally.

Keywords: Unstructured data, Data extraction, Electronic health records, Ophthalmology, Optical character recognition, Natural language processing, Large language model.

1. Introduction

In recent years, the medical information field grows very fast. Now hospitals all over the world use electronic health records. This digital way makes seeing doctors easier, and doctors know the patient better. But it also creates too much clinical data to handle. For the eye department, doctors must do special eye checks and take many pictures. So, the patient data in the eye electronic health record is especially huge.

But there is a problem, most of the most important clinical information in the electronic health record of the eye is not organized in standard format, but in a messy way. These less-standard data are mainly the doctor's handwritten medical records, as well as particularly complex examination pictures such as optical coherence tomography and corneal topographic maps, as well as many of the parameters recorded in the eye examination report [1-3, 13, 25].

Doctors and researchers have long wanted to find useful clinical and scientific information from the mess of unstructured data, mainly by manually looking at and manually recording. This old method makes medical staff particularly tired, it takes a lot of time and energy, when the data is more and more complex, people are easy to see or remember the wrong things [6, 7]. Precision medicine and real-world research are becoming more and more important, and a large amount of high-quality structured data is needed to do a good job of epidemiological investigation, establish disease prediction models, or evaluate the effectiveness of treatment [20, 21]. Therefore, it is now particularly necessary to develop automated data extraction technology, so that the various problems of manual processing can be solved and the key clinical indicators and useful information can be found from the large amount of unstructured eye data faster and more accurate [12, 17].

The data extraction technology development process in the field of ophthalmology can actually be divided into four distinct stages. At first, the technical conditions were limited, and doctors and researchers could only process the paper medical records and examination reports by hand, and re-enter the above information into the computer, which was particularly time-consuming and inefficient. Later, with the development of medical digitization, optical character recognition, that is OCR technology, began to be applied in the field of ophthalmology, this technology mainly solves the problem of turning paper medical records into electronic versions, which can be converted into digital formats that can be read by computers after scanning, which is the first attempt to extract data with automated methods [6, 7, 9].

Now the development of artificial intelligence comes to the third stage, especially after the emergence of deep learning technology, natural language processing, that is, NLP technology began to develop rapidly, in the medical field of text analysis has been a lot of applications. This technology can read the doctor's medical records and surgical records of these less-standard text, from which to find specific disease names, symptoms and various examination values [11, 13, 19], so that more useful medical data can be collected.

Recently, there has been a new change in this field, that is, the large language model LLM began to catch on. The emergence of LLM has made great progress in the automatic extraction of ophthalmic data, because it has a large amount of knowledge and strong understanding ability [26, 27, 30], can accurately find the required information from complex medical records, and can process words and pictures at the same time, and extract important data directly from medical images [25, 28].

In recent years, automatic data extraction techniques such as OCR, NLP and LLM have been used more and more in the electronic medical record system of ophthalmology, which can help doctors process data faster and help make clinical

decisions. However, no one has yet arranged how these technologies are specifically used in the electronic medical records of the eye, and there is no article detailing how these technologies have been developed step by step. This paper is to find out all the literature on this topic and study it again, and make up this part of the work that others have not done.

We divide this paper into a few parts. Part two talks about how we get the right literature. This includes the search way we use and the rules to pick the articles. Part three mainly looks at OCR, NLP, and LLM. These are the three most important auto data extraction tools, and we will see how they do their job. Also, this part puts past literature together to show how these tools actually work on eye electronic medical record data. The last part, part four, will talk about the real problems and hard things when using these auto tools in eye medical records today.

2. Methods

The review wanted to take a comprehensive look at how automated data extraction techniques are now used in the Electronic Health Records (EHRs) for ophthalmology. In order to find all the important literature and ensure that it is of good quality, we have specially designed a relatively strict set of search methods and screening steps.

2.1. Literature Search Strategy

The research used Google Scholar to find papers, as it found many useful articles in different disciplines such as medical informatics, computer science, and ophthalmology. In order to more accurately find the research we need, we designed a more complex search method, which mainly looks at three aspects, one is the data type, one is the technical model, and the clinical application is the third one. Now these technologies are developing particularly fast and are being updated every day, so we don't set too many restrictions on search to better understand how these technologies are developed. Specifically, we used the following keywords to search:

("structured data" OR "unstructured data") AND ("data extraction" OR "information extraction") AND (ophthalmic OR ophthalmology OR ocular OR eye) AND ("electronic health record" OR "electronic health records") AND ("optical character recognition" OR "natural language processing" OR "large language model")

2.2. Literature Screening Criteria

In order to ensure that the quality of the literature we found is good enough, and the content we want to study is particularly relevant, this article has developed some standards to determine which articles can and what not. These criteria are divided into two categories, one of which must be met in order to be selected, and the other is the case in which it cannot be. Only literature that fully meets all requirements will be selected by us, and the specific screening requirements are as follows:

(1) The literature must be peer-reviewed and published in academic journals or conference meetings.

(2) Publication language: English.

(3) The study must apply to the data from eye electronic health records.

(4) The paper must use automated data extraction technology to deal with unstructured data.

(5) Non-review literature.

2.3. Literature Screening Results

By using the search way and keywords we set before, we found 1,610 papers at first. Then we need to check these articles very carefully. First, we delete the repeated ones. After that, we do careful selection for a few rounds. Step one is reading the title and the abstract. If an article is clearly not about eyes or does not use automatic extraction tools, we kick it out. For the articles left, we must read the whole text seriously. Then we use our old rules to decide if we should keep them.

This article has carefully read a lot of articles, and finally selected 30 articles that are particularly well written to do research. These articles mainly use three different technical methods to process data. Ten of them are mainly OCR technology, that is, optical character recognition, which uses this technology to extract useful information from eye pictures and text [1-10]. Another 13 articles focused on NLP techniques, that is, natural language processing [11-23], which mainly used this technique to analyze free text data related to ophthalmology. The remaining articles focus on LLMs, or large language models that perform well in eye data feature extraction and multi-modal parsing [24-30].

3. Results

3.1. Automated Data Extraction Technologies

Eye electronic health records EHRs contain a lot of complex pictures and textual data, these data are not structured, to turn them into structured data that can be used for clinical analysis, mainly rely on three technologies. The first is optical character recognition OCR, which can handle the text in the picture. The second is Natural Language Processing NLP, which is specifically designed to analyze text content. The third is the large language model LLM, which has been used more recently. When these three technologies handle different types of data, because the working principle is not the same, their workflows are also different, and they have their own characteristics.

3.1.1. Optical Character Recognition (OCR)

OCR technology is mainly used to turn those eye instruments or scanned paper generated image reports into text content that can be read by computers. In general, the process of automatically extracting data is like this:

Image Preprocessing: When checking eyes, doctors often use OCT and fundus reports. But these pictures usually have very complicated background things, making them look messy. So we must process the picture first before doing more analysis. We mainly do three steps: change the picture to black and white, clean up the noise we don't want, and fix the contrast so the picture looks sharper. Doing this is just to make the computer recognize it better later [8].

Text Localization and Region of Interest (ROI) Extraction: The main work of the paper focuses on how important data areas can be found from the pictures, like the tables and charts that record the thickness of the retinal nerve fiber layer or the macular thickness. We used two ways to do this, one is the template matching, and the other is the layout analysis algorithm. These ways can help the computer to automatically recognize what is actually useful in the picture, just as the 10th reference stated [10].

Character Recognition and Conversion: The main research in this paper is how to identify the text in the picture into editable text. We did this in two ways [2, 3], old-fashioned

OCR techniques such as Tesseract and the other is deep learning methods that are more popular today. Specifically, deep learning uses models such as CNN and ResNet. The basic principle of these technologies is to divide the picture into a small piece of area [1, 5], and then analyze the pixels of each area, and finally these pixels correspond to the actual text. The whole process is the conversion process from image to text.

Post-processing and Structured Mapping: When the computer gets the original text, we must fix the wrong data. We mostly use specific regular expressions or template rules we made early on to do this. Then, we turn the pulled-out indicators into normal structured data. For example, we change them into normal data files like CSV, or we map them into the common data model CDM. Many studies do this kind of processing, and you can see the related literature [1, 4].

3.1.2. Natural Language Processing (NLP)

NLP technology is mainly used to process the free text of clinical records and surgical reports written by doctors in electronic health records. The process of extracting data automatically probably has several steps:

Text Preprocessing: This step cleans the messy original free text. It does some basic actions like making big and small letters the same, throwing away stop words, cutting words into tokens, and breaking sentences [21].

Feature Extraction and Word Embeddings: In text processing, feature extraction and word embeddings are important methods. In the beginning, people used one-hot encoding to represent the word, and later developed a better way to convert words into low-dimensional continuous vectors. More commonly used are pre-trained eye-specific word embedding models, such as those trained on the PubMed corpus or cui2vec. These models are able to better understand the semantic relationship between ophthalmic technical terms, much better than previous methods [20, 21].

Named Entity Recognition (NER) and Relation Extraction: The paper focuses on the two techniques for named entity recognition and relation extraction, which are particularly important for finding eye symptoms, drug names, and vision values from text. In the early stages of natural language processing, researchers used two methods, one is to check the dictionary according to fixed rules, and the other is to use the Conditional Random Fields (CRF) model in machine learning [15]. However, with the development of deep learning, we began to use some more advanced models [22], such as the bidirectional long short-term memory network and Conditional Random Fields combined with the BiLSTM-CRF model, or directly with the Transformer structure of the pre-trained language model, such as BERT and BioBERT specifically for the biomedical field. These new methods are much more accurate at extracting clinical information than before [13, 17, 19].

Semantic Disambiguation and Standardization Normalization: The main research in this paper is how to deal with the more difficult expression in the ophthalmic medical records, such as the negative meaning of the phrase "no retinal detachment", but also to deal with the time record. We want to sort these contents clearly and then correspond to standard medical terms, such as the CUI number in the UMLS system. This is done to make these medical data more standardized and easier to use. Similar research methods are mentioned in the references [11, 21].

3.1.3. Large Language Models (LLMs)

Now the large language model LLM can be said to be the most powerful technology in the field of generative information extraction, it relies on a large number of parameters and a particularly powerful attention mechanism, completely changed the way we usually do data extraction. In general, the process of automatically extracting data is probably a few steps:

Prompt Engineering: This paper is mainly about prompt engineering, it is not the same as the traditional method, do not need to engage in those particularly complex feature engineering, do not spend a lot of time to do a lot of manual labeling, but by designing some very accurate system instructions to tell the model what to do. In terms of specific operations, researchers often use zero-shot learning this method [30], and sometimes use techniques like few-shot prompting, or specially designed sequential tips for those difficult ophthalmic signs [27].

Multimodal Input and Contextual Parsing: Big models today like GPT-4 and Gemini are very strong. They can read very long and messy medical records. They can even directly understand very hard medical check pictures, like corneal topographic maps and optical biometric reports. They throw all this different information into one system. This helps them understand all kinds of medical data much better [25, 28].

Knowledge Retrieval and Logical Reasoning: In the more difficult task of eye data extraction, such as judging possible complications after surgery, large language models can be used with retrieval augmented generation (RAG) technology. This RAG technology allows the model to find useful information from an external knowledge base, helping it to better understand complex medical situations, so that when logical reasoning is needed, the accuracy of the results is higher [29].

Structured Output Generation: The model can generate things by itself. It can directly spit out structured formats, like JSON files or key-value pairs. This perfectly fits what we need for scientific analysis later. Because of this, the whole extraction job can be finished from start to end [28].

3.1.4. Comparative Analysis of Automated Data Extraction Technologies

When using the three techniques of OCR, NLP and LLM in the electronic medical record system of ophthalmology, each method has its own good and bad place. OCR technology is best used to handle printed, format-fixed reports [1], in which case its accuracy can be almost error-free. However, if you encounter a doctor's handwritten medical record, or the picture quality is poor, or the table and charts are mixed together, OCR's performance will be much worse [4, 6], it is not very adaptable to these complex changes.

NLP works well when analyzing free text written by doctors, such as when it is specifically used to extract the value of vision examination or the name of the drug [13, 18], the accuracy and recall rate are very high. But NLP requires a large number of manually marked data as the basis for training, doctors often use their own habits of abbreviations, different hospital records are not the same [11, 19, 21], these will make the effect of NLP decrease.

In contrast, the advantage of LLM is that it can understand contextual relationships and analyze it with a combination of information, sometimes better than a model designed specifically for a specific field [26], and can be used without special adjustment [29]. LLM can handle particularly complex medical scenarios, but there are still many problems, such as the need for strong computer hardware support, the cost of calling the API is very high, and the privacy and

security of medical data, and sometimes LLM will make up some non-existent values [28, 30].

3.2. Applications of Automated Data Extraction to Clinical Ophthalmology

In the practical use of the electronic health record EHR system for ophthalmology, automated data extraction technology can be applied to many different places. Specifically, when processing different types of unstructured data, such as medical imaging reports, medical record records written by doctors, and clinical data containing multiple forms of information, optical character recognition OCR, natural language processing NLP, and large language model LLM have their own characteristics and can play an important role in clinical work. From the research papers published in recent years, this article mainly wants to clarify how these three techniques specifically help us extract useful data from the eye EHR.

3.2.1. Applications of Optical Character Recognition

In the actual work of the eye hospital, doctors encounter a lot of important patient examination data every day, such as measuring the thickness of the retina, corneal curvature, and the results of the visual field index, but these data are often kept in different inspection instruments, or printed into paper reports. At this time, optical character recognition technology, that is, what we often say OCR, is useful, it can convert the text information on these pictures into text data that can be processed by the computer, and plays a very critical role in the process of organizing these images and text materials.

(1) Data Extraction from Ophthalmic Examination Reports

Optical coherence tomography OCT is particularly important in ophthalmic examinations, and doctors often use it to see the eye condition, especially when diagnosing glaucoma and macular lesions. There are a lot of studies that are using OCR technology to process these OCT reports. For example, Mun has a deep learning OCR system, using ResNet and BiLSTM technology, specifically used to read the RNFL thickness of the retinal nerve fiber layer from OCT images, and found that the accuracy rate can reach 99.67% [1].

In two hospitals, the Tesseract OCR engine was used to process more than 730,000 OCT scans, extracting both the RNFL and the central macular thickness CMT data, and then converting to the standard format of R-CDM, so that the retinal changes of chronically ill patients can be tracked for a long time [2]. Sood also used Tesseract to extract retinal thickness data from high-resolution OCT macular scan reports, and input this data into the data mining algorithm to automatically predict the results of visual field, which is helpful for early detection of glaucoma [3].

In order to extract more accurately in certain regions, Majid used preprocessing algorithms such as image binarization and denoising to specifically process optic nerve OCT reports, so that they can more accurately extract data from the RNFL and the ganglion cell complex (GCC), even if the background is more complicated [8]. In addition, they developed an intelligent document analysis system that can quickly find the places that need attention in the RNFL inspection report, and even the values on the pie chart and radar map can be accurately grasped [10].

(2) Disease Screening Based on Fundus Images or Multimodal Fusion

This paper not only extracts information from numbers, but also uses OCR technology to help deal with image classification. A team of researchers named Luaibi has a powerful system that brings together OCR, image processing and CNN neural networks. The system is very interesting, it can see the problem from the fundus of the eye, and can also extract text information from the doctor's diagnosis report, and finally used to determine whether the patient has diabetic macular edema. The most powerful thing is that the system is particularly accurate, reaching 97% as much [5].

Another researcher called Peissig tried disease screening on a much larger scale. He discovered that putting different methods together works much better. To be specific, he first used OCR to scan the medical record documents. Then he mixed the text information with the billing codes. The result showed that this mixed way is super useful for finding cataract cases. It is much stronger than just using one kind of data, especially when you have to process a huge pile of electronic medical records [9].

(3) Challenges of Paper Forms and Handwritten Medical Records

OCR technology works well when it comes to identifying standard print reports, but when it comes to doctors handwritten notes or old, non-standardized paper forms, it just cannot work. Rasmussen's team tried to build a processing system consisting of multiple OCR modules specifically for the handwritten ophthalmic forms in electronic medical records, although the system is still good in terms of correct identification, but when the doctor's handwriting mark, the recognition rate is only 13.5% so low, which shows that the traditional OCR technology does have a big problem in dealing with free handwriting [6].

In the specific application scenario of myopia screening, the Mendes team used OCR plus machine learning data cleaning methods to process paper versions of children's vision screening records, and found that these paper materials were yellowed and broken because of their age, coupled with various irregular abbreviations written by doctors, the error rate was particularly high at the beginning [7], and finally had to spend a lot of time designing manual verification rules. Kumar's team tested the performance of the OCR models such as AWS Textract, Google Vision and PyTesteract in medical examination reports, and once again proved that although the commercial cloud OCR's recognition accuracy of printed body data can reach 95%, the recognition effect is significantly worse once the handwriting content is encountered [4].

3.2.2. Applications of Natural Language Processing

Now electronic medical records are very common in hospitals, and doctors write a lot of free-form clinical records every day when they visit outpatients or have surgery. Because these records are so good at reflecting patients, researchers are using natural language processing techniques, or NLP, to help find useful clinical information from these complex medical texts.

(1) Extraction of Core Clinical Indicators

Visual acuity VA is particularly important in ophthalmic examinations and is one of the most common indicators of concern to doctors, but the way in which this data is recorded varies greatly in different hospitals, some use Snellen score, some use logMAR, and some use a pinhole. Scientists who first studied this question handled text by writing rules and regular expressions. For example, Mbagwu's team thought of a way to combine the rule algorithm and SQL search, so that the patient's best vision record BDVA can be accurately found from the medical record [15], and the accuracy rate can reach 99%.

An open source tool called TOVA is also very powerful, this tool is Baughman they made, it not only uses regular expressions, but also engages in a scoring system, can read the vision value, but also clear whether the left eye or the right eye test results [23], and experts manually extracted the data basically the same. Wang's team did similar work, using regular expressions with simple NLP classifiers to identify important indicators of cataract and glaucoma surgery from surgical records and review medical records, such as vision recovery and eye pressure changes [12]. But these rule-based models have a big problem, that is, it is not very good to encounter medical records that are not standardized in writing.

To make this trouble go away, Bernstein tried some new ways. They used deep learning models like Transformer. For instance, they used BioBERT and ClinicalBERT which only focus on medical text. The effect of these new models is really good, getting an F1 score of more than 95%. Also, when meeting remote abbreviations or sentences that are very complex, their working performance is much more steady than the old methods [13].

(2) Ophthalmic Cohort Construction and Disease-Specific Feature Extraction

NLP techniques are indeed useful in the medical field, especially in building disease databases and sorting out patient symptoms. Recently, a research team used the AI tool spaCy to identify the name of the disease and other important information in the eye record, and they successfully engaged in an automated eye disease registration system, which saves a lot of money than traditional methods [11]. Another study is also very interesting, they put NLP technology and random forest, support vector machine these machine learning methods, specifically to deal with the discharge records and changes in the condition of hospitalized patients, found that this method can be very accurate to find the case of acute conjunctivitis, but also the patient's symptoms and possible causes are clearly sorted out [14].

(3) Automation of Medication Management and Electronic Case Report Forms

In the field of ophthalmic drug management, Lin and his team did an interesting job using a deep learning model called CNN to identify drug information. This model is designed to deal with outpatient records of patients with glaucoma, to accurately identify the drugs the patient is using, including the name, dosage and frequency of the drug, and to determine whether the patient is taking the drug on time. This technology is very helpful for the hospital to automatically manage patient medication [18].

In addition, the Chen team also made a practical tool, they engaged in an algorithm called Philter, this algorithm combined some rules and white list blacklist methods, you can remove the privacy information in the glaucoma medical record. The advantage of this is that researchers can securely share this medical record data, and they eventually built an open source database of glaucoma drugs [16]. The Wang group also came up with a way to speed up clinical research, and they combined these techniques of BERT, BiLSTM and CRF to develop an NLP processing system. The system can extract all kinds of information before and after the operation directly from the admission record of femtosecond laser surgery to automatically fill in the spreadsheet. After using this method, the time to organize the data is more than 80% less than the original, and the efficiency is improved a lot [17].

(4) Complex Context Comprehension and Multimodal Prognostic Prediction

In the clinical semantic understanding area which is more complex, the group of Wang made a try to use a pre-trained language model called BlueBERT, and added weak supervision to it. They did this to pick out 24 specific eye checks from text with no structure, like the common slit-lamp exam and the fundus exam [19]. But they found out a problem themselves. If they just use the model directly on handwritten medical records that are completely unfixed, the recall rate will turn out to be not good enough.

To overcome the limitations of using only one method, several teams, Gui and Wang, later did some improvement studies, and they used neural word embeddings and autoencoder techniques to combine structured billing data with medical record text obtained through natural language processing. The deep learning predictive model built after this improvement works well, and can predict how likely patients with poor vision will improve in the next year [20, 21].

For the processing of Chinese ophthalmic electronic medical records, the Mao team specially studied the special grammatical characteristics of the history of present illness, they compared the Conditional Random Fields and the Hidden Markov Model of these two methods, and finally found that when Conditional Random Fields are used to do the named entity identification of Chinese ophthalmic symptoms, it is better whether it is accurate or the recall [22].

3.2.3. Applications of Large Language Models

In recent years, large language models, which we often say, have developed particularly fast, and this technological breakthrough has entered a new stage of automated data extraction. Today's LLM is very different from before, they have a lot of parameters, do not need much training to complete a lot of complex reasoning tasks. These models improve the level of text processing a lot, but also deal with different types of data such as pictures and sounds, and even do complex reasoning tasks.

(1) Zero-Shot Information Extraction

Now when doing the task of free text information extraction, the general large model is particularly powerful, basically the strongest. Majid's team experimented with large commercial models like GPT-4 and Gemini, compared to small models trained specifically for a field, such as BioBERT. It was found that when GPT-4 found the name of the drug and the frequency of the drug in the ophthalmic medical record, the F1 score reached 0.962, which was stronger than the best NLP model at that time, and it did not need to do special training in the field of ophthalmology [26]. But there is now a problem, and the big business models that deal with medical data may leak privacy because the data is going to go abroad.

So some researchers began to look at large open source models. These models can be put in the local computers to use. For example, Ruan used the Llama 3.1 model put in local, and they added a set of sequential prompt methods designed by themselves. They did this to accurately find the specific results of slit-lamp inspection and fundus inspection from the text that is particularly messy [27]. Satheakeerthy did this work more thoroughly. They just directly used the Llama-3 model with 70 billion parameters, and they did not do any other training. They extracted the best corrected vision and eye pressure data from the outpatient records, and the accuracy rate reached more than 90 percent. This thing shows that these large open source models are really useful when we do not have labeled data [30].

(2) Multimodal Imaging Report Parsing

This paper focuses on the advantages of multimodal large

language models in medical image report processing. Unlike previous OCR recognition methods, which are not the same as previous OCR recognition and then NLP processing, MLLM can directly understand complex medical image reports in a unified framework. For example, Tan's team has done experiments to feed the Pentacam corneal topographic map report directly to ChatGPT, whether it is a healthy cornea or an abnormal corneal image after surgery, the model can accurately identify all important data, such as corneal curvature, astigmatism and corneal thickness, the accuracy rate reached 100%. Even more, it can also give very reliable diagnostic recommendations based on these values [25].

In addition, Salvi compared the effects of the multi-modal large model Gemini Pro Vision and the traditional OCR tool PaddleOCR processing of the PDF report generated by the optical biometer, and found that the large model accurately captures key parameters of cataract surgery, such as the length of the eye axis and the depth of the anterior chamber, and automatically organizes the structured data that can be directly used by machine learning [28].

(3) Fully Automated Chart Review and Virtual Patient Generation

LLM performs particularly well in complex clinical analysis tasks because of its long background window and strong logically related capabilities. Chen's team made an AI tool called OphthoACR, which combines the adjusted GPT-4 and RAG frameworks to fully automate the review of eye medical charts. The system can not only find the various parameters before the operation, but also accurately find complex post-operative complications in a long and many visits, such as macular edema, so that the original manual processing of a complex medical record took 25 minutes, and now it takes only 80 seconds to use this system [29].

In terms of clinical teaching, Luo used the LLM and RAG framework to transform the electronic medical records of ophthalmic diseases that remove personal information into "Large Language Model Digital Patients", which can have a natural dialogue with medical students. This innovative approach has brought back the role of electronic medical record data that is not usually used, and has greatly improved the ability of medical students to consult digital patients through free dialogue, opening up a new direction for automatic data extraction in medical simulation teaching [24].

4. Discussion

Now the electronic health record of the eye in the use of automated data extraction technology is indeed very promising, this technology allows ophthalmic big data in clinical and scientific research has played a great role, but really to use it in the actual work, or will encounter a lot of technical problems and other difficulties.

4.1. Current Problems and Challenges

4.1.1. Data Access and Privacy Issues

Medical data is particularly sensitive and involves a lot of personal privacy issues, and privacy regulations like HIPAA make it difficult to share it between different hospitals. Everyone is using automated technology to process medical data, but doing so makes privacy protection more difficult. The first problem is that when natural language processing technology automatically removes personal information from text, too much is often deleted. For example, there is a professional term in ophthalmology called "Ahmed glaucoma valve", because there is a personal name in it, the system will delete it as a real name [16], so that important medical information is missing.

The second problem is that people are now using large language models, such as GPT-4 and Gemini, which are mostly closed-source models developed by commercial companies. Hospitals are going to send patients' electronic medical records to the cloud to make these models work, which can easily violate regulations and lead to data breaches [25, 26, 29]. It is precisely because there are so many privacy problems that most studies can only use one hospital's own data, and it is difficult to put together the eye data of many different hospitals.

4.1.2. Data Quality Issues

Eye electronic health record (EHR) data often encounter data loss, noise and inconsistent formatting, which can affect the accuracy of the computer's automatic extraction of data. Take optical character recognition (OCR) technology, those old paper medical records because of the long time there will be wear and stains, the recognition error rate has been very high [7]. Especially when you encounter a doctor's handwritten medical records or fill out forms, the performance of traditional OCR is even worse, and the study found that it can only recognize 13.5% of the handwritten content, which is equivalent to saying that most of the handwritten records cannot be recognized [4, 6].

There are similar problems in natural language processing (NLP), doctors are usually busy at work, writing the progress notes will use their own accustomed abbreviations, but also often misspelling and complex sentence structure [12, 13, 21]. Another big challenge is that real-world medical data itself is incomplete, such as the lack of important laboratory test results, which makes it difficult for computers to make accurate clinical predictions or analyze the cause, no matter how advanced the data extraction technology is used [14].

4.1.3. Inadequate Model Interpretability and Generalization

Many hospitals now use automatic data to extract models, but a big problem is that when these models are used in different hospitals or systems, the effect often becomes very bad. The previous OCR technology has a disadvantage, it must work according to a fixed picture template. If the eye device manufacturer updates the software interface of the inspection report, or changes the chart position, the previous set of extraction rules are completely useless [8].

There is a similar situation in natural language processing, because most of the medical corpus used to train the model is not specifically for ophthalmology [21], and models trained in the general corpus such as cui2vec have encountered more than half of the technical terms of ophthalmic disease. Some studies have also found that language models trained with a hospital's medical record template, such as BERT, directly to another hospital [19], the effect will be particularly poor, because each hospital writes a medical record in a different way.

What's more troubling is that the commonly used deep learning model today is like a black box [13], and doctors simply don't understand how it makes judgments. Where medical care has a particularly low fault tolerance rate, it is hard for doctors to believe the abnormal results of a model that they don't understand.

4.2. Future Research Directions

Now the field of ophthalmology encountered these difficulties, we especially need to find ways to improve the

automation of data extraction technology, not only to innovate in the algorithm, but also to make more efforts in the specific application methods, so as to really solve the problem.

In the field of optical character recognition, that is, we often say OCR technology, the future direction of research may have to change the way of thinking. Now many methods rely too much on fixed templates, which is actually not very good. We feel that we can combine the two technologies of object detection and document layout analysis to make a more intelligent visual model. Especially for something that changes often in the structure of eye instrument diagrams, this new method may be more useful. Some studies also support this idea, indicating that this direction is worth trying [10].

In the natural language processing or NLP area, future studies should not just stare at the text words. They must also look at the pictures and tables. Specifically, we need to analyze the chart data read by OCR tools, the text information pulled out by NLP, and the neat clinical data. Doing this helps us fully understand the patient's real situation from many sides. Past research shows that combining data in this multi-faceted way really helps us deeply grasp the patient's whole condition [9, 20].

Large language model, which we often call LLM, is indeed promising in terms of automatic data processing, but this model still has many problems to be solved in the practical application of ophthalmic electronic medical records.

Reducing and Suppressing "Hallucinations": This article discusses how to solve the "hallucinations" problem that occurs in the medical data of generative models. When medical records are incomplete, these models may concoct some data to fill in the gaps, which can cause a lot of problems in practical applications. For example, when analyzing optical biometric reports, large language models occasionally replace missing data items with values of other indicators [28], as mentioned in the literature. In order to solve this problem, the next study can consider adding a stricter confidence scoring system, and combine RAG technology to verify the accuracy of the facts [29], so that each result of the model output can find the corresponding original text basis.

Enhancing Multimodal Fusion and Complex Logical Reasoning: Eye exams mostly rely on many kinds of imaging techniques. So the big language models in the future will do better in this aspect. To be specific, these models need to analyze retinal OCT images, corneal topographic maps, and the medical records written by doctors at the same time. They need to put them in the same system to understand these different types of information [25, 28]. Now some patients maybe have to do several surgeries or go for follow-up checks in a very short time. When this situation happens, the model often messes up the time order. It also mistakes some normal conditions as the complications after surgery. To solve this problem [29], the models in the future should focus on improving the ability to process long-term medical records. By doing this, they can track the changes in the condition more accurately.

Constructing Lightweight, Ophthalmology-Specific Large Models: Now to do a lightweight model for ophthalmology, but directly take the ready-made model to do zero-shot inference effect is particularly poor, even the doctor wrote in the medical records of those do not understand the place can not be clear [30], such as confusing the left and right eyes or did not write clearly before or after correction of vision data. If you use those open-source super-large models, such as the 70 billion parameter Llama 3 [27, 30], which runs locally and is too much hardware, the general hospital simply can't afford it. Therefore, the next focus is to study how to fine-tune such technologies with parameter-efficient fine-tuning, and retrain the open source model on the eye professional data. This is a lightweight eye-specific model that can understand both the professional content written by the doctor and the local deployment in private hospitals, which not only ensures the accuracy of the medical record, but also protects the patient's private data.

5. Conclusion

The use of automated data extraction technology in ophthalmic electronic health records has become very important, and it helps us to find more useful clinical and scientific information from these records. This review mainly talks about how this field has been developed step by step, starting with optical character recognition OCR technology, which solves the problem of conversion between pictures and text [1, 2], so that those standardized eye instrument reports and previously recorded paper data can be digitized.

Later, the natural language processing NLP technology that appeared was even more powerful, and it could read the free-form medical records written by doctors, whether it was vision inspection results [13, 21], eye pressure data or various complex eye disease characteristics, which can be accurately identified. The combination of these technologies has revolutionized the entire process of data extraction, from simple records at the beginning to complex diagnostic reports of later images [25, 29, 30].

At present, the use of automated extraction technology in the field of ophthalmology does face many problems, such as how to protect the privacy of patients, deal with those data that are not uniform in a format, and let the model adapt to more different situations. However, these technologies are still quite large in the application space of ophthalmology. With the development of technology, especially the more mature methods that can better integrate different types of data together, coupled with the lightweight large language models designed for ophthalmology that can be used directly in local hospitals, these advances will improve the research ability in real medicine. In this way, ophthalmology can be supported by more reliable data in precision medicine and intelligent assisted diagnosis.

References

[1] Mun Y, Kim J, Noh K J, et al. An innovative strategy for standardized, structured, and interoperable results in ophthalmic examinations [J]. BMC Medical Informatics and Decision Making, 2021, 21(1): 9.

[2] Park C H, Lee S H, Lee D Y, et al. Analysis of Retinal Thickness in Patients With Chronic Diseases Using Standardized Optical Coherence Tomography Data: Database Study Based on the Radiology Common Data Model [J]. JMIR Medical Informatics, 2025, 13: e64422.

[3] using Structural A G D. Automated Glaucoma Detection using Structural Optical Coherence Tomography With Data Mining [J].

[4] Kumar A, Singh P, Lata K. Comparative study of different optical character recognition models on handwritten and printed medical reports [C]//2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). IEEE, 2023: 581-586.

[5] Luaibi Z H, Alkhayyat A N. Detection and Analysis of Diabetic Macular Edema (DME) Using Artificial Intelligence Techniques [J]. Journal of Madenat Al-Elem University College/Magalla ẗ Kulliyya ẗ Madīna ẗ Al- ʿ Alam Al- Ğ ā mi'a ẗ, 2025, 17(2).

[6] Rasmussen L V, Peissig P L, McCarty C A, et al. Development of an optical character recognition pipeline for handwritten form fields from an electronic health record [J]. Journal of the American Medical Informatics Association, 2012, 19(e1): e90-e95.

[7] Mendes I, Miranda V, Salazar M, et al. Enhancing Amblyopia Screening with Machine Learning: Challenges and Solutions in Data Preparation [J]. Procedia Computer Science, 2025, 257: 1092-1097.

[8] Majid I, Zhang Y V, Chang R, et al. Extraction of Text from Optic Nerve Optical Coherence Tomography Reports [J]. arXiv preprint arXiv:2308.10790, 2023.

[9] Peissig P L, Rasmussen L V, Berg R L, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records [J]. Journal of the American Medical Informatics Association, 2012, 19(2): 225-234.

[10] Hua C, Shi Y, Hu M, et al. Intelligent data extraction system for RNFL examination reports [C]//CAAI International Conference on Artificial Intelligence. Cham: Springer Nature Switzerland, 2022: 537-542.

[11] Macri C Z, Teoh S C, Bacchi S, et al. A case study in applying artificial intelligence-based named entity recognition to develop an automated ophthalmic disease registry [J]. Graefe's Archive for Clinical and Experimental Ophthalmology, 2023, 261(11): 3335-3344.

[12] Wang S Y, Pershing S, Tran E, et al. Automated extraction of ophthalmic surgery outcomes from the electronic health record [J]. International journal of medical informatics, 2020, 133: 104007.

[13] Bernstein I A, Koornwinder A, Hwang H H, et al. Automated recognition of visual acuity measurements in ophthalmology clinical notes using deep learning [J]. Ophthalmology science, 2024, 4(2): 100371.

[14] Valdes Sanz N, Garcia-Layana A, Colas T, et al. Clinical characterization of inpatients with acute conjunctivitis: a retrospective analysis by natural language processing and machine learning [J]. Applied Sciences, 2022, 12(23): 12352.

[15] Mbagwu M, French D D, Gill M, et al. Creation of an accurate algorithm to detect Snellen best documented visual acuity from ophthalmology electronic health record notes [J]. JMIR medical informatics, 2016, 4(2): e14.

[16] Chen J S, Lin W C, Yang S, et al. Development of an open-source annotated glaucoma medication dataset from clinical notes in the electronic health record [J]. Translational Vision Science & Technology, 2022, 11(11): 20-20.

[17] Wang B, Lai J, Liu M, et al. Electronic source data transcription for electronic case report forms in China: validation of the electronic source record tool in a real-world ophthalmology study [J]. JMIR Formative Research, 2022, 6(12): e43229.

[18] Lin W C, Chen J S, Kaluzny J, et al. Extraction of active medications and adherence using natural language processing for glaucoma patients [C]//AMIA Annual Symposium Proceedings. 2022, 2021: 773.

[19] Wang S Y, Huang J, Hwang H, et al. Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam [J]. International journal of medical informatics, 2022, 167: 104864.

[20] Wang S Y, Singh S, Njie Jr S. Looking for Low Vision in Electronic Health Records [J].

[21] Gui H, Tseng B, Hu W, et al. Looking for low vision: predicting visual prognosis by fusing structured and free-text data from electronic health records [J]. International journal of medical informatics, 2022, 159: 104678.

[22] Mao X, Li F, Duan Y, et al. Named entity recognition of electronic medical record in ophthalmology based on crf model[C]//2017 International conference on computer technology, electronics and communication (ICCTEC). IEEE, 2017: 785-788.

[23] Baughman D M, Su G L, Tsui I, et al. Validation of the total visual acuity extraction algorithm (TOVA) for automated extraction of visual acuity data from free text, unstructured clinical records [J]. Translational Vision Science & Technology, 2017, 6(2): 2-2.

[24] Luo M J, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills [J]. NPJ Digital Medicine, 2025, 8(1): 502.

[25] Tan J C K, Coroneo M T. Diagnostic interpretation of corneal tomography using a multimodal large language model (ChatGPT) [J]. American Journal of Ophthalmology Case Reports, 2025: 102441.

[26] Majid I, Mishra V, Ravindranath R, et al. Evaluating the performance of large language models for named entity recognition in ophthalmology clinical free-text notes [C]//AMIA Annual Symposium Proceedings. 2025, 2024: 778.

[27] Ruan F Y, Lam J W, Esmaeilkhanian H, et al. Leveraging Large Language Models with Sequential Prompting to Extract Eye Examination Findings from Free-Text Ophthalmology Notes [J]. Ophthalmology Science, 2025: 100944.

[28] Salvi A, Arnal L, Ly K, et al. Ocular Biometry OCR: a machine learning algorithm leveraging optical character recognition to extract intra ocular lens biometry measurements [J]. Frontiers in Artificial Intelligence, 2025, 7: 1428716.

[29] Chen K M, Chen K W, Diaconita V, et al. OphthoACR (Ophthalmology Automated Chart Review): An AI-Powered Tool for Complete Automation of Ophthalmology Chart Reviews and Cohort Data Analysis [J]. Translational Vision Science & Technology, 2025, 14(10): 8-8.

[30] Satheakeerthy S, Jesudason D, Bahrami B, et al. Zero-shot LLM-based visual acuity extraction: a pilot study [J]. BMC ophthalmology, 2025, 25(1): 359.