中山大學 中山眼科中心 ZHONGSHAN OPHTHALMIC CENTER, SUN YAT-SEN UNIVERSITY

Contents lists available at Lighthouse Scholar

EYE SCIENCE

journal homepage: journal.gzzoc. com

doi: 10.12419/es24030101
View this article at: https://dx.doi.org/10.12419/es24030101

· Original Article ·

Harnessing AI-human synergy for deep learning research analysis in ophthalmology with large language models assisting humans

Mingjie Luo(罗明杰)[1§], Weixing Zhang(张玮星)[1§], Zheming Zhang(张哲铭)[1], Jianyu Pang(庞健宇)[1], Zhenzhe Lin(林桢哲)[1], Lanqin Zhao(赵兰琴)[1], Duoru Lin(林铎儒)[1], Haotian Lin(林浩添)[1,2,3]

1. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou 510060, China
2. Center for Precision Medicine and Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510060, China
3. Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou 570311, China

HIGHLIGHTS

· Introduction of an AI-human collaboration framework tailored for comprehensively assessment of trends in the medical research field.
· In-depth analysis of trends in ocular diseases, image modalities, and research data quality and volume using large language models assisting humans.
· Insightful revelations on the historical advances and future directions of deep learning in ophthalmology.

Abstract: **Background:** Research innovations inocular disease screening, diagnosis, and management have been boosted by deep learning (DL) in the last decade. To assess historical research trends and current advances, we conducted an artificial intelligence (AI)–human hybrid analysis of publications on DL in ophthalmology. **Methods:** All DL-related articles in ophthalmology, which were published between 2012 and 2022 from Web of Science, were included. 500 high-impact articles annotated with key research information were used to fine-tune a large language models (LLM) for reviewing medical literature and extracting information. After verifying the LLM's accuracy in extracting diseases and imaging modalities, we analyzed trend of DL in ophthalmology with 2 535 articles. **Results:** Researchers using LLM for literature analysis were 70% (P = 0.000 1) faster than those

Received date: 2024-02-15; **Accepted date:** 2024-03-20; **Published online:** 2024-03-28

§Contributing equally

Corresponding authors: Haotian Lin, E-mail: linht5@mail.sysu.edu.cn. Duoru Lin, E-mail: lindr5@mail.sysu.edu.cn.

Full Text

who did not, while achieving comparable accuracy (97% versus 98%, P = 0.768 1). The field of DL in ophthalmology has grown 116% annually, paralleling trends of the broader DL domain. The publications focused mainly on diabetic retinopathy (P = 0.000 3), glaucoma (P = 0.001 1), and age-related macular diseases (P = 0.000 1) using retinal fundus photographs (FP, P = 0.001 5) and optical coherence tomography (OCT, P = 0.000 1). DL studies utilizing multimodal images have been growing, with FP and OCT combined being the most frequent. Among the 500 high-impact articles, laboratory studies constituted the majority at 65.3%. Notably, a discernible decline in model accuracy was observed when categorizing by study design, notwithstanding its statistical insignificance. Furthermore, 43 publicly available ocular image datasets were summarized. **Conclusion:** This study has characterized the landscape of publications on DL in ophthalmology, by identifying the trends and breakthroughs among research topics and the fast-growing areas. This study provides an efficient framework for combined AI–human analysis to comprehensively assess the current status and future trends in the field.

Keywords: large language model; AI–human collaboration; research trends; ophthalmology; model performance

Cite this article as: Luo MJ, Zhang WX, Zhang ZM, Pang JY, Lin ZZ, Zhao LQ, Lin DR, Lin HT. Harnessing AI–human Synergy for Deep Learning Research Analysis in Ophthalmology with Large Language Models Assisting Humans. Eye Science, 2024, 1(1): 7-25. doi: 10.12419/es24030101

INTRODUCTION

There has been a significant advancement in the field of biomedical artificial intelligence (AI) since 2012, particularly in the medical applications of deep learning (DL) algorithms. Medical AI plays an increasingly important role in the development of medicine and improvement of the health and longevity of the population. This period witnessed the transition from the maturity of DL algorithms to their incorporation into various medical domains, marking a complete phase for biomedical AI development. Ophthalmology has experienced significant growth and development in the field of DL, with both early-stage and well-established DL applications being utilized in real-world scenarios.[1] With the technical breakthrough of DL algorithms in the last decade, the field have gradually achieved intelligent diagnosis of different ocular diseases such as diabetic retinopathy (DR),[2-3] cataract[4] and glaucoma,[5] and have expanded the coverage of different image modalities from retinal photograph to optical coherence tomography (OCT) and other modalities. Therefore further clarification and refinement are needed to fully comprehend the developmental patterns and trends in the field of DL in ophthalmology.

Recent publications on DL in ophthalmology has shown explosive growth. These publications provide wealth of valuable information about the changing trends in DL-assisted diagnostic systems for ocular diseases, along with a comparative analysis of the development status in different countries worldwide. This field acts as an excellent demonstration of the evolution of medical AI, showcasing changes in various aspects and providing insights into future trends. Nonetheless, published literature reviews or expert consensuses have largely focused on specific research areas or highly cited articles,[6–9] limiting their scope and providing only a narrow perspective on the overall changes in the field of DL in ophthalmology. Consequently, comprehensive insights into the broader dimensions of ophthalmic AI

have been difficult to obtain, such as the trend of research in ocular diseases, changes in data modalities, quantities and quality of research data, along with the factors affecting the DL model performance.[10] Additionally, manually reading and analyzing through such a large volume of literature is a time-consuming and challenging task, and might lead to a biased representation of the overall perspective.[11] It is urgently needed to conduct a comprehensive and practical overview of this fast-evolving field.

Recently, large language models (LLM) have shown significant success in following instructions and producing human-like responses.[12-13] Using LLM for natural language processing and applying the state-of-the-art bibliometric analysis in joint with human experts, this study presented a comprehensive overview of the evolution in this rapid changing research field. Base on this, we have identified trends and challenges among common ophthalmic DL research and further provided prospects for future applications. Additionally, this study offers a practical approach to comprehensively investigate current status and future trends in the field, making it a valuable reference for other researchers.

MATERIALS AND METHODS

Search Strategy and Data Collection

All publications related to DL in ophthalmology in the English language in the Web of Science (WOS) database were queried. The search included terms related to eye diseases, eye structures and common imaging examinations in ophthalmology, as well as DL-related terms such as convolutional neural networks and transfer learning. The detailed search terms used are included in the supplementary text (Supplementary text). The literature search was carried out on October 15, 2022, and studies published from January 1, 2012, to September 30, 2022 were included. A total of 6,345 articles were initially identified, 3,810 articles among them were excluded for the following reasons: duplicated records, articles categorized as reviews and comments, and articles tangential to the core focus of ophthalmology. 2,535 articles among them met the inclusion criteria after screening (Figure 1A). Among the 2,535 articles, 1,260 articles were indexed by the PubMed database. To utilize the rich information in the PubMed database, we obtained additional metadata such as the MeSH terms for these articles. The metadata were downloaded and extracted using PubMed's E-Utilities API tools (www.pubmed.ncbi.nlm.nih.gov/help). The article title, abstract, MeSH terms and bibliometrics, including WOS citation numbers, countries, regions and institutes, were used for downstream publication analysis.

AI-human Hybrid Publication Analysis

These articles were then analyzed using the LLM-based text analysis method (Figure 1B). Out of the 2,535 articles, the top 500 most impactful articles were selected for manual in-depth analysis based on their WOS citation counts on October 30th, 2022. Three individual researchers independently annotated key information in the articles, including disease, study design, data modality, data quality and quantity and public datasets. The annotated data were then used to fine-tune the bioBERT model,[14] and the tuned LLM was then used to automatically recognize disease and data modality information in the rest of articles. Additionally, BERN2,[15] a validated disease recognition model, was used to verify the disease recognition outputs from both models. In the LLM–assisted comparison experiment, 20 articles were randomly selected from the 2,535 articles. Three researchers were then assigned to read the titles and abstracts, extract information described above, and record the time spent and accuracy achieved.

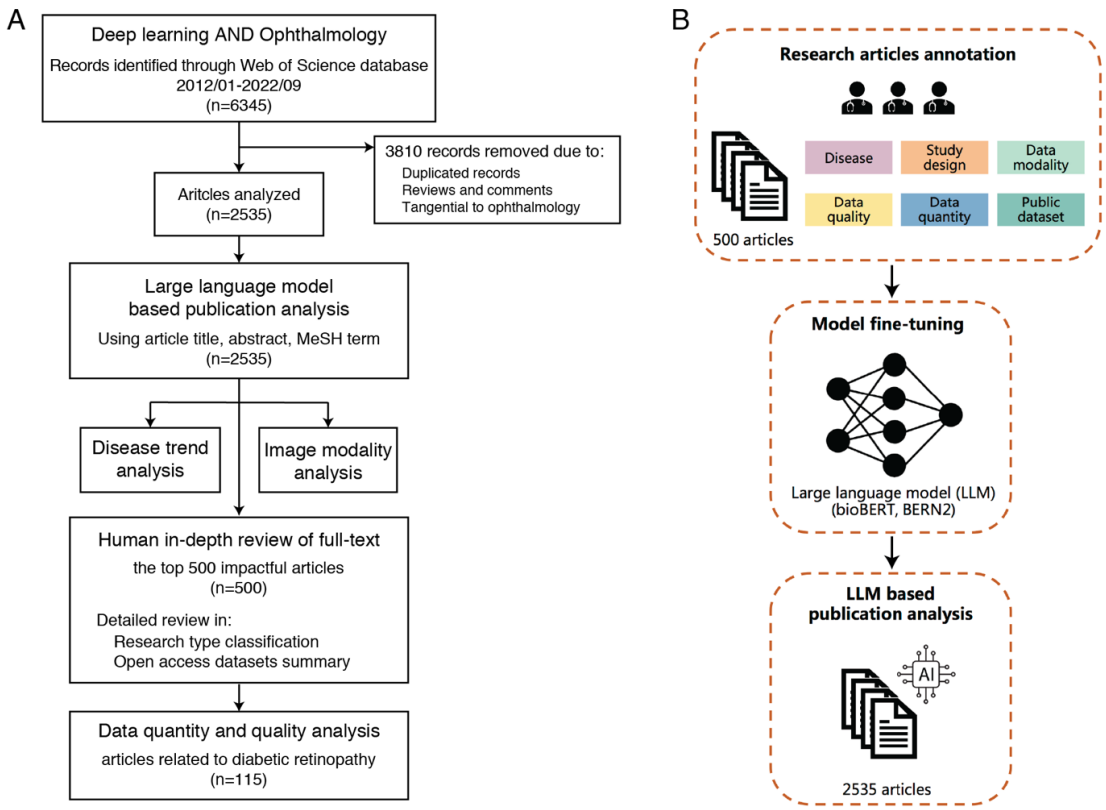

Figure 1 Overall flow diagram of this study.

(A) Flow diagram summarizing the study screening and analysis procedures. A total of 6,345 articles were initially identified by querying the Web of Science database. After screening 2,535 articles met the inclusion criteria and were analyzed using an LLM-based text analysis method. The top 500 most impactful articles were included for human in-depth analysis. (B) Overview of the research article annotation process and the model fine-tuning procedure. LLM: large language model; MeSH: medical subject headings.

Using the LLM-based text analysis method, we successfully extracted the studied eye diseases from 2,300 of 2,535 (90.7%) articles in the field of DL ophthalmology and categorized numerous disease types based on PubMed's disease MeSH terms, and the data modalities used were extracted and summarized from 1,611 of 2,535 (63.5%) articles. Assessing the quality and quantity of study data was relatively challenging, and required manual analysis given the limitations of LLMs in mathematical processing. To ensure accuracy, the data quality and quantity analysis included results from manual review of the 500 highest impact articles. In addition, 115 DR-related DL studies were further selected to analyze the dataset used including whether the studies had external validation datasets, used public datasets, a dataset with images of healthy controls and provided pixel-level annotations (identifying the boundaries of disease lesions).

Research Type Classification

The articles were classified into three categories: laboratory, preclinical and clinical research. Laboratory studies involve constructing and validating algorithms and models using public data. Clinical studies utilize prevalidated AI algorithms and models in real clinical scenarios to assist in disease diagnosis and intervention. Preclinical studies, a type between laboratory and clinical studies, focus on the algorithm or model validation through public or limited private datasets to assess their future suitability for large-scale clinical use. Preclinical

and clinical research studies were classified into three categories according to their design: retrospective, cross-sectional, and prospective. Study types were initially classified by two junior researchers (>3 years research experience). In cases of disagreement, adjudication was performed by a senior researcher (>8 years research experience) to finalize the study type.

Public Ophthalmic Image Datasets

Original papers were retrieved corresponding to the public ophthalmic image datasets used in the 500 highest-impact papers. Then, we compiled and summarized the following information: database name, article DOI\URL, year of publication, type of disease, number of images, number of healthy controls, image disease annotations, image quality assessment, and pixel-level annotations.

Statistical Analysis

All statistical analyses in this study were conducted with R 4.1.1 (R Core Team, 2021). Research trends were analyzed with Sen's slope analysis. The normality of the distribution of data quantity was tested with the Shapiro-Wilk test. Linear regression was employed to analyze the correlations between research publication date and data quantity. The model performances were compared with the Mann-Whitney U test. A two-tailed p-value less than 0.05 was considered statistically significant for all analyses.

RESULTS

Efficiency Improvement of the LLM-assistant Approach

To evaluate the efficiency of using an LLM as an auxiliary tool in literature reading and information extraction, two groups of researchers, each consisting of three individuals, were tasked with extracting information from a set of 20 research articles (Supplementary Table S1). The group utilizing the LLM completed the task in an average time of 39 (range 35–45) minutes, while the group without the LLM took an average of 128 (119–137) minutes, yielding a 70% ($P = 0.0001$) increase in efficiency. The accuracies of the two groups were comparable, 97% (95%–100%) versus 98% (95%-100%, $P = 0.7681$).

Overview of DL Applications in Ophthalmology

A total of 2,535 articles were included in the LLM–assisted analysis. Over the past decade, both papers describing DL applications in ophthalmology and those describing DL in general have been witnessed a dramatic surge in publication volumes. Ophthalmic DL articles experienced explosive growth rates between 2012 and 2019, with yearly growth rates ranging from 111% to 192%. This growth rate was moderate after 2020, with 659 articles published in 2021; moreover, this trend of DL applications in ophthalmology parallels the changes in the overall DL field (Figure 2A). There were several technology breakthroughs related to the growth of the ophthalmic DL field (Figure 2B). The U.S., China, and the U.K. published the most articles, with the former two countries leading by a considerable margin (Figure 2C). Additionally, the top ten institutions with the highest publication numbers mainly comprised comprehensive universities and eye hospitals (Figure 2D).

Diseases and Data Modalities in Ophthalmic DL Research

The 10 most commonly studied diseases were shown in Figure 3A, including DR, glaucoma, macular degeneration, cataract and fundus diseases. As stated here, researchers related to various ocular diseases began to gradually increase in 2016 and showed a significant increase between 2017 and 2020. The most studied

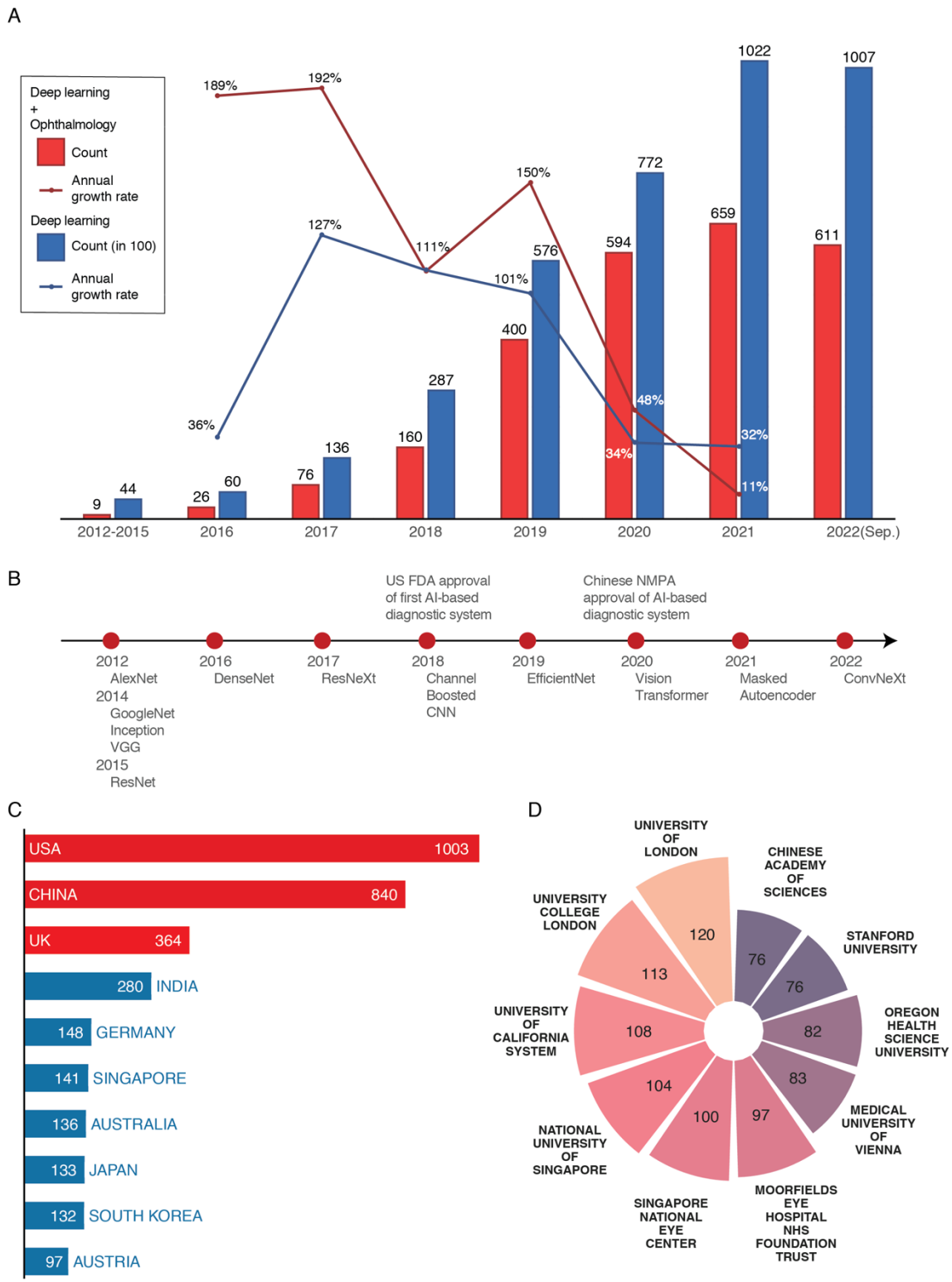

Figure 2 The number and growth rate of publications per year in the field of deep learning in ophthalmology and milestone events in computer vision.

(A) The number of articles published per year in the fields of deep learning (blue, numbers shown in hundreds) and DL-related ophthalmology (red). (B) Timeline of significant deep learning models since 2012. Countries (C) and institutions (D) with the largest number of ophthalmology studies related to deep learning.

on ocular diseases are DR, glaucoma and macular degeneration, which have also shown an increasing trend in the last 10 years, with Sen's slopes of 20.8 (P = 0.0003), 14.9 (P = 0.0011) and 8.2 (P = 0.0001), respectively (Figure 3B).

With similar methods, the data modalities used were extracted and summarized. The top 10 data modalities in terms of growth rate were ranked on Sen's slope (Figure

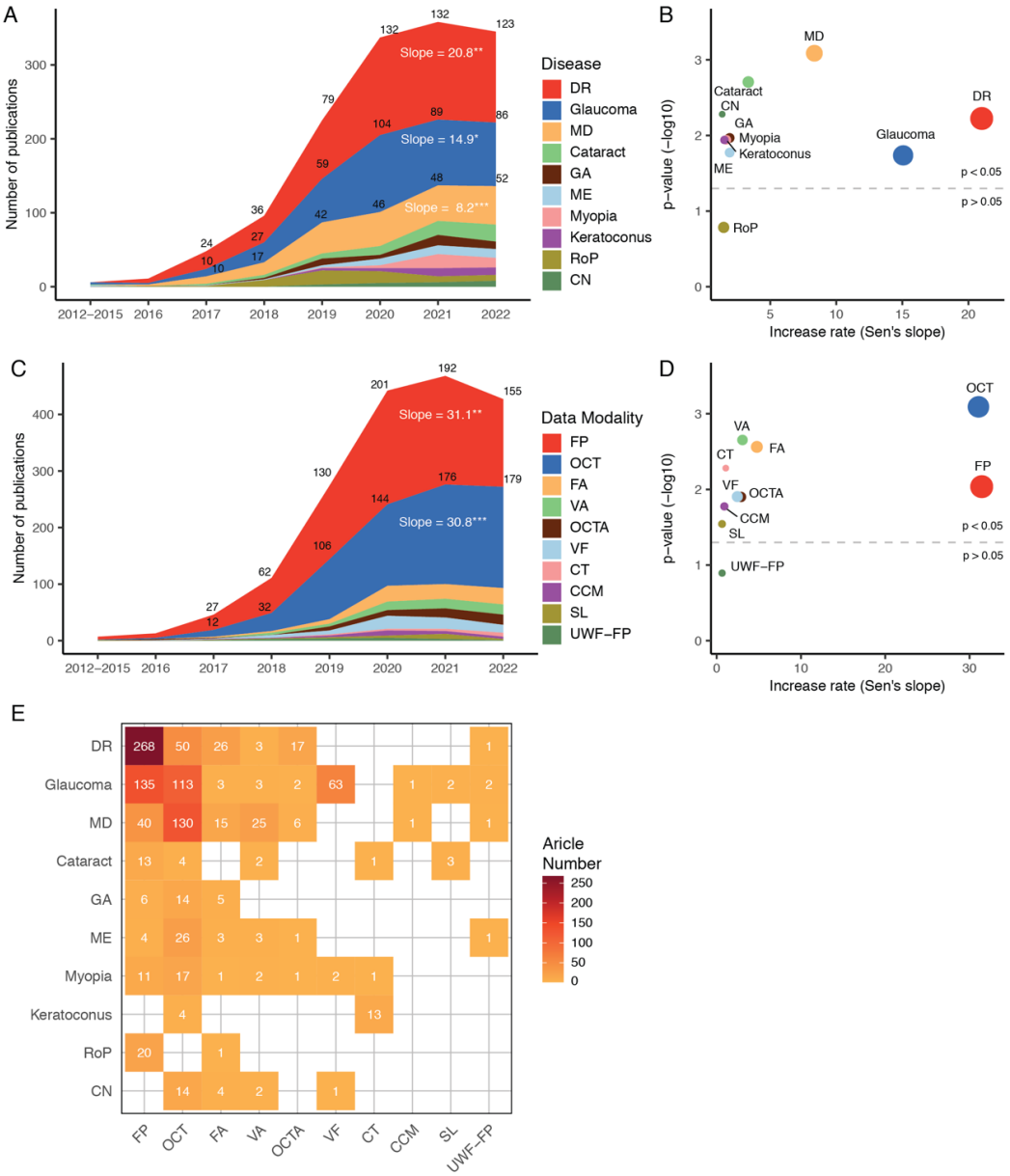

Figure 3 Historical trends of the ocular diseases considered and data modalities used in deep learning studies. The number of published articles per year for the top 10 diseases.

(A) and data modalities (C) with the largest increases. Increase rate analysis for different types of ocular diseases (B) and data modalities (D). Sen's slope represents the increase rate and –log10(p value) indicates significance. Sen's slopes are shown for diseases: DR, glaucoma, MD and modalities: OCT and FP. (E) A heatmap showing the cooccurrence of ocular diseases and related data modalities. FP: fundus photographs; OCT: optical coherence tomography; FA: fluorescein angiography; FAF: fundus autofluorescence; VF: visual fields; SL: slit lamp; VA: visual acuity; OCTA: optical coherence tomography angiography; CT: corneal topography; CCM: corneal confocal microscopy; UWF-FP: ultrawide-field retinal fundus photographs; DR: diabetic retinopathy; MD: macular degeneration; GA: geographic atrophy; ME: macular edema; RoP: retinopathy of prematurity; CN: choroidal neovascularization.

3C). As stated, DL studies using fundus photographs (FP) and OCT began to rapidly increase in number and gain prominence in 2018, with Sen's slopes of 31.1 (P = 0.0015) and 30.8 (P = 0.0001), respectively (Figure 3D). Many other data modalities including fluorescein angiography (FA), optical coherence tomography angiography (OCTA), visual acuity (VA), visual fields (VF) and corneal topography (CT), have been increasing applied since 2020. The data modalities used in studies of different ocular diseases varied considerably (Figure 3E). Specifically, studies on DR mainly used OCT and FP; studies on glaucoma relied mainly on OCT, FP, and VF; and macular degeneration studies relied on OCT and FP. Additionally, several DL studies used multimodal ophthalmic data (Table 1). 27 (33%) multimodality studies involved both FP and OCT images, 17 (21%) used OCT and VF, and 7(9%) studies utilized FP and one of the modalities such as OCTA, slit lamp (SL), and VA.

Research Type Analysis of the Top 500 Highest-impact Articles

Among the top 500 most impactful articles, all three types of research (laboratory, preclinical and clinical) were found a rise in number of studies (Figure 4A). The number of laboratory research studies increased noticeably from 7 in 2012 to 108 in 2020. Furthermore, preclinical studies experienced a discernible rise in number to over 50 per year, whereas the number of clinical studies remained stagnant, with fewer than 10 conducted annually. Among preclinical and clinical studies, retrospective studies accounted for 77.4% (123), with only 21 (12.6%) cross-sectional studies, and 15 (10.1%) prospective studies (Figure 4B).

Changes in Fundus Photograph Data and DL Model Performance for DR

Given the extensive researches on DR and its early prominence, out of 500, 115 DR-related articles were selected to evaluate changes in quantity and quality for ophthalmic data and DL model performance. The dataset characteristics of these studies were summarized in Table 2. All criteria showed a general increase since 2015, despite occasional volatility. Although the early dearth of available datasets and use of healthy control images, the

Table 1 Summary of multimodal deep learning studies in ophthalmology

Multimodalities	Number of studies	Study ID[1]
FP + OCT	27 (33%)	45; 46; 48; 50; 51; 52; 56; 57; 58; 59; 61; 66; 67; 68; 71; 72; 73; 74; 75; 76; 77; 79; 80; 81; 92; 108; 111
FP + FA	8 (10%)	45; 52; 53; 54; 64; 70; 78; 82
FP + OCTA/SL/VF/ UWF/VA	7 (9%)	52; 47; 49; 114; 128; 120; 63
OCT + VF	17 (21%)	60; 65; 69; 91; 97; 99; 100; 101; 104; 107; 110; 116; 117; 119; 121; 122; 129
OCT + OCTA	12 (15%)	90; 94; 96; 98; 102; 103; 105; 106; 113; 118; 124; 127
OCT + FA/ICGA/ FAF	7 (9%)	43; 55; 62; 93; 95; 123; 125
OCT + VA/UBM	3 (4%)	109; 112; 126

[1] Each study ID corresponds to a particular research article as listed in Supplemental table S1. FP: fundus photographs; OCT: optical coherence tomography; FA: fluorescein angiography; FAF: fundus autofluorescence; ICGA: indocyanine green angiography; VF: visual fields; SL: slit lamp; UBM: ultrasound biomicroscopy; VA: visual acuity.

number of studies fulfilling these criteria have increased to 14 (33.3%) and 25 (59.5%), respectively. Additionally, an increasing number of studies tended to provide pixel-level lesion annotations over bulk image annotations, showing an increase in the data granularity.

The dataset size and final model performance used in the 115 DR studies were compiled. This is to determine whether a large data size is necessary to train a model with adequate performance (Figure 4C). The interquartile range (IQR) of the data size for DR classification models is 462–74,198 images, and the average data size has grown from 70,817 images in 2016 to 122,810 in 2021, with an average annual increase of 10,398 (14.7%) images. By dividing the presented performances in these studies according to the study design, it was revealed that the average accuracy of the models declined consistently (Figure 4D).

To provide a reference of publicly available datasets, the ocular disease image databases used in these 500 highest-impact articles were summarized with the pertinent information (Table 3 and Supplementary Table S2).

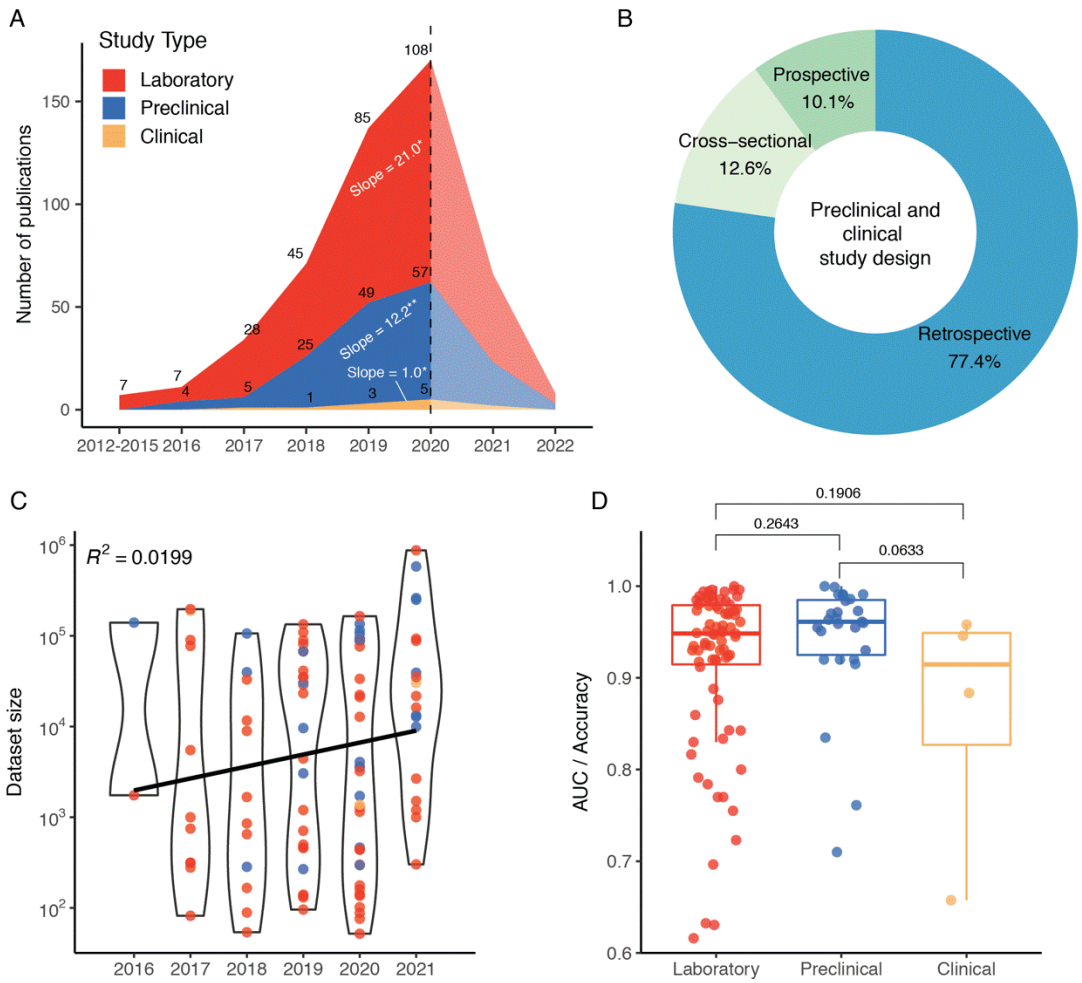

Figure 4 Trend in type of study and the magnitude of data size for ophthalmic DL studies. (A) Trends in ophthalmic DL studies for different research types; (B) The study designs of the 159 preclinical and clinical studies. (C) The data size used in ophthalmic deep learning research increased every year (R^2 = 0.0199, p = 0.1). The data were fitted using linear regression and R^2 and p-values were calculated. The axis for data size is displayed on a base-10 logarithmic scale. (D) Comparison of model performance between 3 study designs. The Mann–Whitney U test was used. AUC: area under the receiver operating characteristic curve.

Table 2 Data characteristics from the 115 most impactful deep learning studies of diabetic retinopathy Data characteristics from the 115 most impactful deep learning studies of diabetic retinopathy

Data characteristics	2016	2017	2018	2019	2020	2021
Number of studies identified	3	13	17	36	42	23
Studies with external validation datasets	1	2	4	3	10	4
	(33.3%)	(15.4%)	(23.5%)	(8.3%)	(23.8%)	(17.4%)
Studies using public datasets	0	3	6	10	14	7
	(0.0%)	(23.1%)	(35.3%)	(27.8%)	(33.3%)	(30.4%)
Studies with images from healthy controls	0	7	10	17	25	16
	(0.0%)	(53.8%)	(58.8%)	(47.2%)	(59.5%)	(69.6%)
Studies with pixel-level lesion annotations	1	5	11	15	22	13
	(33.3%)	(38.5%)	(64.7%)	(41.7%)	(52.4%)	(56.5%)

Table 3 Summary of the open access ophthalmological datasets used in deep learning studies

Image modality	Public datasets[1]	Year	Disease(s)	Number of images
FP	ACRIMA	2019	Glaucoma	705
FP	APTOS 2019	2019	Diabetic retinopathy	5,590
FP	Bin Rushed	2019	Glaucoma	195
FP	CHASE dataset	2011	Retinal arteriolar tortuosity	16,670
FP	Chinese Glaucoma Study Alliance	2019	Glaucoma	274,413
FP	Chiu_BOE	2015	Diabetic macular edema	110
FP	DDR	2019	Diabetic retinopathy	13,673
FP	DESP	2020	Diabetic retinopathy	174,954
FP	DIARETDB0	2007	Diabetic retinopathy	130
FP	DIARETDB1	2007	Diabetic retinopathy	89
FP	Direct-CSU	2019	Glaucoma	934
FP	DRIMDB	2014	Retinal diseases	216
FP	DRIONS-DB	2018	Glaucoma	110
FP	DRISHTI-GS	2018	Glaucoma	101

[1] Detailed characterization information and references for the public datasets were provided in Supplementary Table S1. FP: fundus photographs; OCT: optical coherence tomography; FA: fluorescein angiography; FAF: fundus autofluorescence; VF: visual fields; SL: slit lamp.

DR was predominant in terms of disease type, followed by glaucoma, AMD and fundus vascular diseases. Correspondingly, most ocular databases consist of fundus photographs, followed by those containing OCT images, OCTA, and FA. In addition, the AREDS and SEED databases contain a variety of data modalities: AREDS has FP and SL data, while SEED contains FP, OCT, and SL data.

Table 3 (continued)

Image modality	Public datasets[1]	Year	Disease(s)	Number of images
FP	DRISHTI-GS1	2014	Glaucoma	101
FP	DRIVE	2004	Diabetic retinopathy	400
FP	e-Ophtha	2013	Diabetic retinopathy	25,702
FP	FGADR	2008	Diabetic retinopathy	2,842
FP	HEI-MED	2020	Diabetic retinopathy	1,907
FP	HRF	2013	Diabetic retinopathy and glaucoma	45
FP	IDRiD	2018	Diabetic retinopathy	516
FP	Kaggle	2015	Diabetic retinopathy	88,702
FP	Magrabi	2019	Glaucoma	94
FP	MESSIDOR	2008	Diabetic retinopathy	1,200
FP	NEW SINDI	2018	Glaucoma	5,783
FP	NIH AREDS	2018	Age-related macular degeneration	5,664
FP	ORIGA	2010	Glaucoma	650
FP	REFUGE	2018	Glaucoma	1,200
FP	RIGA	2016	Glaucoma	750
FP	RIM-ONE	2011	Glaucoma	169
FP	STARE	2000	Various retinal diseases	397
FP	UK biobank	2017	Various eye diseases	135,867
FP, SL	AREDS	1999	Age-related macular degeneration, age-related cataract	9,386
FP, OCT, SL	SEED	2021	Various age-related eye diseases	10,033
OCT	Heidelberg-DME	2020	Diabetic retinopathy	1,396
OCT	Triton-DME	2020	Diabetic retinopathy	3,248
OCT	UniMiami	2019	Diabetic retinopathy	50
SD-OCT	A2A SD-OCT	2012	Age-related macular degeneration	345
SD-OCT	SERI-CUHK	2019	Diabetic retinopathy	43
SD-OCT	Duke	2013	Age-related macular degeneration	38,800
AS-OCT	SCES	2012	Anterior chamber depth	1,060
OCTA	ROSE	2020	Retinal disease vessel segmentation	229
FA	RECOVERY-FA19	2020	Retinal disease vessel detection	8
FA	VAMPIRE FA	2020	Diabetic retinopathy	8

DISCUSSION

DL in ophthalmology is evolving rapidly and the status of the field is difficult to summarize comprehensively using conventional methods.[7–9] This study presents a comprehensive overview of all existing research using LLM-based methods combined with in-depth manual analysis to uncover trends in DL in ophthalmology from a large amount of publication data. The development of DL in ophthalmology is in sync with the overall progress in DL domain, relying heavily on the foundational support provided by the latter. In this study, we fine-tuned the BERT-based large-scale language model with manually annotated paper data to build an intelligent LLM capable of reviewing medical literature and extracting key information. Its accuracy was verified in extracting disease and image modality information, suggesting that incorporating an LLM into the literature analysis workflow could enhance the efficiency and productivity of researchers in the field. The number of articles and development trends of various eye diseases were categorized and summarized using the LLM–assisted approach.

Advances in the DL field have accelerated DL research in ophthalmology. From 2012 to 2015, there were revolutionary developments in the DL field, including the construction of benchmark image datasets ImageNet,[16] COCO[17] and development of convolutional neural network architectures such as AlexNet[18] and ResNet.[19] The subsequent rapid growth of DL research in ophthalmology occurred as a result of these breakthroughs in DL advancing the state-of-the-art and laying the foundation for further research and development of DL in ophthalmology. Many of the leading ophthalmic deep learning studies have come from technologically advanced countries including the United States, China, the United Kingdom, and India. With major research contributions from these countries, deep learning-based diagnostic systems for ocular diseases have gained approval from global regulatory bodies. In 2018, the first U.S.license for such a system was issued,[20] followed by the Chinese license in 2020.[21] This widespread regulatory approval reflects the progress in deep learning research and validation of its ability to accurately detect and diagnose eye diseases, signaling a shift from laboratory studies to clinical applications.

Ophthalmic DL research has gradually transitioned from initial feasibility studies to real-world clinical applications. Early feasibility DL studies focused on DR based on FP, as the large patient population and extensive labeled image datasets enabled algorithm development and validation with minimal technical complexity.[22] By establishing capabilities and clinical value on a common ocular disease (e.g. DR) with abundant training data, researchers laid the groundwork to then investigate expanding DL to other ocular diseases.[1] The initial successes have led to further research broadening real-world deployment of DL models across diverse clinical applications, such as telemedicine screening and smartphone-based diagnostic tools at point-of-care.[23-24]

Moving beyond reliance on single data modalities like FP and OCT, there is instead increasing utilization of multimodal approaches that integrate information from diverse clinical exams and tests to enable more comprehensive and accurate diagnosis. This transformation is driven by complex ocular conditions like glaucoma that require assimilating data from basic ophthalmic exams, visual field testing, cup-to-disc ratios from fundus imaging, and optic nerve layer thickness from OCT to facilitate robust clinical judgments.[25] Moreover, combining other data sources like genomic tests with imaging data has been proven efficacy

for predictive modeling in diseases like age-related macular degeneration.[26] As algorithms become more sophisticated, they can synergistically combine disparate inputs from various ophthalmic subspecialties, testing modalities, and data types. By amalgamating these diverse datasets, DL models can mimic multifaceted clinical decision making and enable more precise disease diagnosis and prognosis across ophthalmology.

Data quantity and quality are crucial in DL applications in ophthalmology. However, the necessary sample size is contingent upon the complexity of the disease and detection task, as well as the intricacy of the model. In an early experiment, the impact of dataset size on DL algorithm performance in detecting DR was analyzed.[3] The results revealed that peak performance was reached at approximately 60,000 images, suggesting that increasing the dataset size beyond this point did not improve algorithm performance. However, with advancements in DL algorithms, it is now unclear what amount of data is required for optimal performance. By investigating high-impact DL articles on DR detection, our findings indicate that the DL model's performance decreased markedly from laboratory to clinical studies. This observation implies that the self-reported AUC and other evaluation criteria employed in these studies may not adequately represent the real-world performance of the DL models due to the reproducibility issues.[27] Alternatively, it could be that the data analyzed, extracted from previously published DR-related studies, lack broad applicability to other ocular diseases, which necessitates more rigorous experimental investigations in future studies. Additionally, we found that the proportion of studies using externally validated datasets and public datasets was not high (20%-35%) which might be due to the difficulty in obtaining resources for external validation datasets and public datasets. The question about the sample size required for deep learning is the one without a standard answer. Determining the optimal sample size for DL studies is challenging due to the disease complexity, specific medical tasks, and the complexity of DL models.[28-29] Given that the results derived from different deep learning algorithms on various datasets cannot be directly compared, it is imperative to establish a unified and objective standard evaluation method. This will ensure a greater consistency in the assessment of model performance, thereby enhancing the reliability of the outcomes.

This study has several limitations. This study is based on previously published articles, and therefore may not capture emerging trends in research that has not yet been published. Additionally, the citation time-frame used in this study is restricted to high-impact articles published mainly before 2021, resulting in fewer articles from 2022 onward that were included in the refined analysis. To obtain more concrete conclusions, a more robust study with a larger sample size is warranted. Although we analyzed ophthalmology research trends, data modality, volume, and types of studies, some other aspects of deep learning in ophthalmology research were not addressed in this paper. These aspects include data privacy and security, interpretability and transparency of AI models, and regulation and standardization of AI in practical applications. There are some articles that are not indexed in PubMed and do not have MeSH words, which is mitigated by our thorough analysis of titles and abstracts, ensuring a comprehensive review that minimizes the impact of this limitation on our study's findings.

In conclusion, we showed that an LLM combined with in-depth manual analysis were capable of reviewing medical literature and extracting information. Using the LLM–assisted approach, we have identified trends and challenges among common ophthalmic DL research

and further provided prospects for future applications. This includes the necessity of validating AI models via real world clinical setting, and creating standardized, public accessible datasets to enhance collaboration, benchmarking for DL applications. Additionally, this study offers a practical approach to comprehensively investigate current status and future trends in the field, making it a valuable reference for other researchers.

Correction notice

None

Acknowledgement

None

Author Contributions

(I) Conception and design: HTL, DRL, MJL, ZZL

(II) Administrative support: HTL, DRL

(III) Provision of study materials or patients: HTL, DRL

(IV) Collection and assembly of data: MJL, WXZ, ZMZ, JYP

(V) Data analysis and interpretation: MJL, JYP, ZZL, LQZ

(VI) Manuscript writing: MJL, WXZ

(VII) Final approval of manuscript: All authors

Fundings

This study was supported by the National Natural Science Foundation of China (82000946), Guangdong Natural Science Funds for Distinguished Young Scholar (2023B1515020100), the Natural Science Foundation of Guangdong Province (2021A1515012238), and the Science and Technology Program of Guangzhou (202201020522 and 2022010 20337).

Conflict of Interests

None of the authors has any conflicts of interest to disclose. All authors have declared in the completed the ICMJE uniform disclosure form.

Patient consent for publication

None

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Provenance and Peer Review

This article was a standard submission to our journal. The article has undergone peer review with our anonymous review system.

Data Sharing Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res. 2019; 72: 100759. DOI: 10.1016/j.preteyeres.2019.04.003.
2 Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017; 318: 2211–2223. DOI: 10.1001/jama.2017.18152.

3 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016; 316: 2402–2410. DOI: 10.1001/jama.2016. 17216.

4 Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng. 2017; 1: 0024. DOI: 10.1038/s41551-016-0024.

5 Dong L, He W, Zhang R, et al. Artificial intelligence for screening of multiple retinal and optic nerve diseases. JAMA Netw Open. 2022; 5: e229960. DOI: 10.1001/jamanetworkopen.2022.9960.

6 Sebastian A, Elharrouss O, Al-Maadeed S, Almaadeed N. A survey on deep-learning-based diabetic retinopathy classification. Diagnostics. 2023; 13: 345. DOI: 10.3390/diagnostics13030345.

7 Yang J, Wu S, Dai R, et al. Publication trends of artificial intelligence in retina in 10 years: Where do we stand? Front Med. 2022; 9. DOI:10.3389/fmed.2022. 1001673.

8 Lim WX, Chen Z, Ahmed A. The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review. Med Biol Eng Comput. 2022; 60: 633–642. DOI: 10.1007/s11517-021-02487-8.

9 Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives. Adv Ophthalmol Pract Res. 2022; 2: 100078. DOI: 10.1016/j.aopr.2022.100078.

10 Münchmeyer J, Woollam J, Rietbrock A, et al. Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. J GEOPHYS RES-SOL EA. 2022; 127: e2021JB023499.

11 Lee CJ, Sugimoto CR, Zhang G, et al. Bias in peer review. J AM SOC INF SCI TEC. 2013; 64: 2–17.

12 Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023; 614: 214–216.

13 Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023. DOI: 10.1001/jama.2023.1044.

14 Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36: 1234–1240. DOI: 10.1093/bioinformatics/btz682.

15 Sung M, Jeong M, Choi Y, et al. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics. 2022; 38: 4837–4839. DOI: 10.1093/bioinformatics/btac598.

16 Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis. 2015; 115: 211–252. DOI: 10.1007/s11263-015-0816-y.

17 Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. Computer Vision–ECCV 2014. Cham: Springer International Publishing; 2014: 740–755. DOI: 10.1007/978-3-319-10602-1_48.

18 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2012. DOI:10.1145/ 3065386.

19 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE; 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

20 U.S. Food and Drug Administration (FDA). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye; 2018 Accessed June 13, 2023.

21 National Medical Products Administration (NMPA). Diabetic retinopathy fundus image assisted diagnosis software product approved for marketing. https://www.nmpa.gov.cn/yaowen/ypjgyw/20200810093435157.html; 2020 Accessed June 13, 2023.

22 Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018; 1: 39. DOI: 10.1038/s41746-018-0040-6.

23 Natarajan S, Jain A, Krishnan R, et al. Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone. JAMA Ophthalmol. 2019; 137: 1182–1188. DOI: 10.1001/jamaophthalmol.2019.2923.

24 Lin D, Xiong J, Liu C, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. Lancet Digit Health. 2021; 3: e486–e495. DOI: 10.1016/S2589-7500(21)00086-8.

25 Li F, Su Y, Lin F, et al. A deep-learning system predicts glaucoma incidence and progression using retinal photographs. J Clin Invest. 2022, 132(11): e157968. DOI: 10.1172/JCI157968.

26 Yan Q, Weeks DE, Xin H, et al. Deep-learning-based prediction of late age-related macular degeneration progression. Nat Mach Intell. 2020, 2(2): 141-150. DOI: 10.1038/s42256-020-0154-9.

27 Chen B, Wen M, Shi Y, et al. Towards training reproducible deep learning models. Proceedings of the 44th International Conference on Software Engineering.; 2022: 2202–2214. DOI: 10.1145/3510003.3510163.

28 Rajput D, Wang W-J, Chen C-C. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics. 2023; 24: 48. DOI: 10.1186/s12859-023-05156-9.

29 Figueroa RL, Zeng-Treitler Q, Kandula S, et al. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012; 12: 8. DOI: 10.1186/1472-6947-12-8.

Supplementary Text

A. The full set of search keywords was based on 3 parts:

1. Keywords related to deep learning and computer vision:

(TS="deep learning" OR TS="convolutional neural network" OR TS="transfer learning")

2. Keywords related to ophthalmology:

2.1 Ocular disease names

TS="diabetic retinopathy" OR TS="cataract" OR TS="congenital cataract" OR TS="glaucoma" OR TS="blindness" OR TS="age-related macular degeneration" OR TS="retinal disease" OR TS="diabetic macular edema" OR TS="vision loss" OR TS="retinopathy" OR TS="drusen" OR TS="keratoconus" OR TS="retinopathy of prematurity" OR TS="myopia" OR TS="eye disease" OR TS="visual impairment" OR TS="retinopathy of prematurity" OR TS="choroidal neovascularization" OR TS="geographic atrophy" OR TS="macular edema" OR TS="vision impairment" OR TS="retinal disorders" OR TS="ocular disease" OR TS="macular hole" OR TS="papilledema" OR TS="meibomian gland dysfunction" OR TS="retinal detachment" OR TS="dry eye"

2.2 Ocular structure names

TS="retina" OR TS="fundus" OR TS="cornea" OR TS="choroid" OR TS="vitreous" OR TS="sclera" OR TS="fovea"

2.3 Common ocular examinations

TS="fundus image" OR TS="optical coherence tomography image" OR TS="OCT image" OR TS="slit lamp image" OR

TS="ultrasound biomicroscopy image" OR TS="fluorescein angiography" OR TS="fundus autofluorescence" OR TS="indocyanine green angiography" OR TS="near-infrared reflectance" OR TS="red free"

3. Article publication period of January 2012 – September 2022

DOP=(2012-01-01/2022-09-30)

Combined search keywords

(TS="deep learning" OR TS="convolutional neural network" OR TS="transfer learning") AND (TS="diabetic retinopathy" OR TS="cataract" OR TS="congenital cataract" OR TS="glaucoma" OR TS="blindness" OR TS="age-related macular degeneration" OR TS="retinal disease" OR TS="diabetic macular edema" OR TS="vision loss" OR TS="retinopathy" OR TS="drusen" OR TS="keratoconus" OR TS="retinopathy of prematurity" OR TS="myopia" OR TS="eye disease" OR TS="visual impairment" OR TS="retinopathy of prematurity" OR TS="choroidal neovascularization" OR TS="geographic atrophy" OR TS="macular edema" OR TS="vision impairment" OR TS="retinal disorders" OR TS="ocular disease" OR TS="macular hole" OR TS="papilledema" OR TS="meibomian gland dysfunction" OR TS="retinal detachment" OR TS="dry eye" OR TS="retina" OR TS="fundus" OR TS="cornea" OR TS="choroid" OR TS="vitreous" OR TS="sclera" OR TS="fovea" OR TS="fundus image" OR TS="optical coherence tomography image" OR TS="OCT image" OR TS="slit lamp image" OR TS="ultrasound biomicroscopy image" OR TS="fluorescein angiography" OR TS="fundus autofluorescence" OR TS="indocyanine green angiography" OR TS="near-infrared reflectance" OR TS="red free") AND DOP=(2012-01-01/2022-09-30)

B. Search processes

1. A WOS search using the above search keywords (WOS link: https://www.webofscience.com/wos/alldb/summary/b6a98fd9-b2a4-44cc-8ad6-4d274bcfd99e-5a146c4d/relevance/1)

2. A search in which the "field of research" is limited to ophthalmology (WOS link: https://www.webofscience.com/wos/alldb/summary/18b67078-701e-4f66-9af3-533aa35f9a94-5a15082a/relevance/1)

Supplementary Table S1 Comparison of time and accuracy between researchers working with and without assistance from the LLM

Items		LLMM-assisted group	Non LLM-assisted group
Time(minutes)	Researcher 1	45	137
	Researcher 2	38	130
	Researcher 3	35	119
	Mean	39	128
Accuracy(%)	Researcher 1	95	95
	Researcher 2	97	99
	Researcher 3	100	100
	Mean	97	98

Supplementary Table S2 Comprehensive summary of the open-access ophthalmological datasets used in deep learning studies

Image modality	Public datasets	Year	Disease(s)	Number of images	Images labels	Mutiple expert annotation	Pixel-level lesion annotation	Images from healthy controls	Reference
FP	ACRIMA	2019	Glaucoma	705	Y	N	N	Y	doi: 10.1186/s12938-019-0649-y
FP	APTOS 2019	2019	Diabetic retinopathy	5,590	Y	N	N	Y	link: www.kaggle.com/c/aptos2019-blindness-detection
FP	Bin Rushed	2019	Glaucoma	195	Y	N	N	N	doi 10.1007/s10792-016-0329-x
FP	CHASE dataset	2011	Retinal arteriolar tortuosity	16,670	Y	N	N	Y	doi: 10.1161/atvbaha.111.225219
FP	Chinese Glaucoma Study Alliance	2019	Glaucoma	274,413	Y	N	N	Y	doi: 10.1001/jamaophthalmol.2019.3501
FP	Chiu_BOE	2015	Diabetic macular edema	110	Y	N	N	Y	doi: 10.1364/BOE.6.001172
FP	DDR	2019	Diabetic retinopathy	13,673	Y	N	N	Y	doi: 10.1016/j.ins.2019.06.011
FP	DESP	2020	Diabetic retinopathy	174,954	N	N	N	Y	doi: 10.1136/bjophthalmol-2020-316594
FP	DIARETDB0	2007	Diabetic retinopathy	130	Y	N	N	N	link: www.it.lut.fi/project/imageret/diaretdb0
FP	DIARETDB1	2007	Diabetic retinopathy	89	Y	N	N	Y	link: www.it.lut.fi/project/imageret
FP	Direct-CSU	2019	Glaucoma	934	Y	N	N	Y	doi: 10.1109/JBHI.2019.2934477
FP	DRIMDB	2014	Retinal diseases	216	Y	N	N	N	doi: 10.1117/1.JBO.19.4.046006
FP	DRIONS-DB	2018	Glaucoma	110	N	N	N	Y	doi: 10.1016/j.eswa.2018.06.010
FP	DRISHTI-GS	2014	Glaucoma	101	N	N	N	Y	doi: 10.1109/ISBI.2014.6867807
FP	DRIVE	2004	Diabetic retinopathy	400	Y	N	N	N	link: www.drive.grand-challenge.org/
FP	e-Ophtha	2013	Diabetic retinopathy	25,702	Y	N	N	Y	doi: org10.1016/j.irbm.2013.01.010
FP	FGADR	2020	Diabetic retinopathy	2,842	Y	N	N	Y	doi: 10.1109/TMI.2020.3037771
FP	HEI-MED	2020	Diabetic retinopathy	1,907	Y	N	N	Y	doi: 10.1016/j.cmpb.2020.105398
FP	HRF	2013	Diabetic retinopathy and glaucoma	45	N	N	N	N	link: www.cs.fau.de/research/data/fundus-images
FP	IDRiD	2018	Diabetic retinopathy	516	Y	N	N	Y	doi: 10.3390/data3030025
FP	Kaggle	2015	Diabetic retinopathy	88,702	N	N	N	N	link: www.kaggle.com/c/diabetic-retinopathy-detection

Supplementary Table S2 (continued)

Image modality	Public datasets	Year	Disease(s)	Number of images	Images labels	Mutiple expert annotation	Pixel-level lesion annotation	Images from healthy controls	Reference
FP	Magrabi	2019	Glaucoma	94	Y	N	N	N	doi: 10.1109/CBMS.2019.00100
FP	MESSIDOR	2014	Diabetic retinopathy	1,200	N	N	N	Y	doi: 10.5566/ias.1155
FP	NEW SINDI	2018	Glaucoma	5,783	N	N	N	N	doi: 10.1109/TMI.2018.2837012
FP	NIH AREDS	2018	Age-related macular degeneration	5,664	Y	N	N	N	doi: 10.1016/j.compbiomed.2017.01.018
FP	ORIGA	2010	Glaucoma	650	Y	N	N	Y	doi: 10.1109/IEMBS.2010.5626137
FP	REFUGE	2018	Glaucoma	1,200	Y	N	N	Y	link: www.refuge.grand-challenge.org
FP	RIGA	2016	Glaucoma	750	Y	N	N	Y	doi: 10.1007/s10792-016-0329-x
FP	RIM-ONE	2011	Glaucoma	169	Y	N	N	Y	doi: 10.1109/cbms.2011.5999143
FP	STARE	2000	Various retinal diseases	397	Y	N	N	N	http://www.ces.clemson.edu/~ahoover/stare
FP	UK biobank	2017	Various eye diseases	135,867	N	N	N	N	link: www.ukbiobankeyeconsortium.org.uk
FP, SLP	AREDS	1999	Age-related macular degeneration, cataract	9,386	N	N	N	N	doi: 10.1016/s0197-2456(99)00031-8
FP, OCT, SLP	SEED	2021	Various age-related eye diseases	10,033	Y	N	N	N	doi: 10.1093/ije/dyaa238
OCT	Heidelberg-DME	2020	Diabetic retinopathy	1,396	Y	N	N	Y	doi: 10.1109/JBHI.2020.2983730
OCT	Triton-DME	2020	Diabetic retinopathy	3,248	Y	N	N	Y	doi: 10.1109/JBHI.2020.2983730
OCT	university of miami	2019	Diabetic retinopathy	5	Y	N	N	Y	doi: 10.1002/jbio.201500239
SD-OCT	A2A SD-OCT	2012	Age-related macular degeneration	345	Y	N	N	Y	doi:10.1016/j.ophtha.2012.07.004
SD-OCT	CUHK	2019	Diabetic retinopathy	43	Y	N	N	Y	doi: 10.1109/ICSIPA.2017.8120661
SD-OCT	Duke	2013	Age-related macular degeneration	38,800	Y	Y	N	N	doi: 10.1016/j.ophtha.2013.07.013
AS-OCT	SCES	2012	Anterior chamber depth	1,060	N	Y	N	N	doi: 10.1016/j.ophtha.2012.01.011
OCTA	ROSE	2020	Retinal vessel segmentation	229	Y	N	N	Y	doi: 10.1109/TMI.2020.3042802
FA	RECOVERY-FA19	2019	Retinal vessel detection	8	Y	N	N	Y	doi: 10.21227/m9yw-xs04
FA	VAMPIRE FA	2020	Diabetic retinopathy	8	Y	N	N	N	doi: 10.1109/JBHI.2020.2999257