https://doi.org/10.1007/s44196-023-00225-6

REVIEW ARTICLE

# A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions

Xiangdong Pei[1,2] · Ke Zuo[1] · Yuan Li[1] · Zhengbin Pang[1]

Received: 12 May 2022 / Accepted: 16 March 2023

## Abstract

In recent years, deep learning has been applied in the field of clinical medicine to process large-scale medical images, for large-scale data screening, and in the diagnosis and efficacy evaluation of various major diseases. Multi-modal medical data fusion based on deep learning can effectively extract and integrate characteristic information of different modes, improve clinical applicability in diagnosis and medical evaluation, and provide quantitative analysis, real-time monitoring, and treatment planning. This study investigates the performance of existing multi-modal fusion pre-training algorithms and medical multi-modal fusion methods and compares their key characteristics, such as supported medical data, diseases, target samples, and implementation performance. Additionally, we present the main challenges and goals of the latest trends in multi-modal medical convergence. To provide a clearer perspective on new trends, we also analyzed relevant papers on the Web of Science. We obtain some meaningful results based on the annual development trends, country, institution, and journal-level research, highly cited papers, and research directions. Finally, we perform co-authorship analysis, co-citation analysis, co-occurrence analysis, and bibliographic coupling analysis using the VOSviewer software.

**Keywords** Multi-modal deep learning · Medical · Fusion · Bibliometrics

## Abbreviations

| | |
|---|---|
| MMDLM | Multi-modal deep learning in medicine |
| MRI | Magic resonance imaging |
| PET | Positron emission computed tomography |
| EHR | Electronic health record |
| CT | Computed tomography |
| ConVIRT | Contrastive learning of medical visual representations from paired images and text |
| DALL·E | A neural network model for multi-modal pre-training |
| CLIP | Contrastive language–image pre-training |
| BLIP | Bootstrapping language-image pre-training |
| CogView | A neural network model for multi-modal pre-training |
| WenLan | Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training |
| UniT | Unified Transformer |
| UNITER | Universal image-text representation |
| ViLT | Vision and language Transformer |
| CPt | Colorful prompt tuning |
| ALBEF | Align before fuse |
| ITC | Image-text contrast learning |
| MLM | Masking language modeling |
| ITM | Image-text matching |
| AUC | Area under the receiver operating characteristic curve |
| CNN | Convolutional neural network |
| GCN | Graph convolutional network |
| TCGA | The Cancer Genome Atlas |
| RNN | Recurrent neural network |
| GPT-3 | Generative pre-trained Transformer-3 |
| LUSC | Lung squamous cell carcinoma |
| GBM | Glioblastoma multiforme |
| BRCA | Breast invasive carcinoma |
| RNA | Ribonucleic acid |
| LGG | Brain lower grade glioma |
| ADNI | Alzheimer's Disease Neuroimaging Initiative |
| DETR | DEtection TRansformer |
| BERT | Bidirectional encoder representation from Transformers |
| DOF | Deep orthogonal fusion |

✉ Zhengbin Pang
zbpang@nudt.edu.cn

[1] College of Computer, National University of Defense Technology, Changsha 410073, China

[2] Shanxi Supercomputing Center, Lvliang 033000, China

Published online: 29 March 2023

| | |
|---|---|
| DCE | Dynamic contrast-enhanced |
| BiLSTM | Bidirectional LSTM |
| LSTM | Long short-term memory |
| MoCo | Momentum contrast |
| InfoNCE | Info noise contrastive estimation |
| HCPs | Highly cited papers |
| RCNN | Regions with CNNs |
| WSI | Whole slide image |
| H&E | Hematoxylin and eosin |

## 1 Introduction

### 1.1 Background

The concept of multi-modal systems is related to the study of information representation in the field of human–computer interaction. The term “mode” refers to the representation and exchange of information on a specific physical medium. Due to the development of medical technology and science, medical image fusion has gained wide attention in the field of image processing. Medical imaging methods can be broadly divided into two types: anatomical imaging and functional imaging. Single-mode medical images can only provide a particular aspect of health information and cannot fully reflect all the information in certain parts of the body [1]. In clinical practice, doctors often need to comprehensively analyze different types of medical images of the same part to diagnose the patient's condition; this increases the difficulty of diagnosis. Therefore, fusion processing of multi-modal medical images can be used to comprehensively analyze different medical image information in a fusion image and provide doctors with a more adequate basis for the judgment of clinical diagnosis and treatment [2].

Deep learning provides scientific methods for processing large-scale medical images and screening big data, as well as for the diagnosis and efficacy evaluation of various major diseases in clinical medicine. This major scientific problem in medical image analysis needs to be solved urgently, considering it is a key cutting-edge medical image technology [1]. Multi-modal medical image fusion based on deep learning can be used to effectively extract and integrate the feature information of different modes, improve the clinical applicability of medical images in the diagnosis and evaluation of medical problems, and provide quantitative analysis, real-time monitoring, and treatment planning for doctors and researchers. Multi-modal fusion combines the information of multiple modes for target prediction (classification or regression), which was previously understood as multi-source information fusion [3]. For example, videos as a type of multimedia can be subdivided into multiple single modes, such as dynamic text, dynamic images, and dynamic voice [4]. Research shows that information processing methods based on the multi-modal concept often perform better than traditional single-model methods [5].

Multi-modal medical image fusion is the process of fusing multiple images using single or multiple imaging methods to improve the image quality while preserving specific features [6]. Medical image fusion involves several interesting fields, such as image processing, computer vision, pattern recognition, machine learning, and artificial intelligence, and hence has been widely used in clinical practice. Doctors can understand lesions in different ways through the application of medical image fusion.

### 1.2 Motivation for this Paper

Multi-modal deep learning is the fusion of various types of information via deep-learning techniques. Imaging technology plays an important role in medical diagnosis. The information provided by a single-mode medical image is limited since large amounts of information need to be processed in clinical diagnosis. In multi-modal technology, a single mode of the medical image can supplement the weakness of another mode to accurately evaluate the medical condition and obtain diagnostic information through the fusion of information from multiple modes. Furthermore, multi-modal deep learning can jointly learn the potentially shared information of each mode data through the complementary fusion of different feature sets, which improves the effectiveness and accuracy of data tasks [7]. Additionally, multi-modal medical image fusion based on deep learning can effectively extract and integrate the feature information of different modes, thereby improving the clinical applicability of medical images in the diagnosis and evaluation of medical problems. Considering the above advantages, the application of multi-modal deep learning in medicine (MMDLM) has rapidly attracted wide attention.

This study was conducted to address the following gaps in the existing literature:

1. a lack of effective literature on the multi-modal data fusion mechanisms of medical images to sort and summarize research points.
2. the limited knowledge of data analysis algorithms as well as a lack of clinical experts and data scientists.
3. insufficient information on the bibliometric analysis of the application of multi-modal deep learning in medical imaging.

Based on the above-mentioned points, we conducted a comprehensive review and bibliometric analysis of this field to explore potential models or scientific development paths of multi-modal deep learning in medical image applications.

The contributions of our study are as follows:

1. We macroscopically analyze the scope of medical multi-modal applications and mainstream pre-training methods, and discuss the adaptability of each method.
2. We discuss various topics in medical multi-modal fusion methods, ranging from micro-scale methods (such as algorithm-based convolutional neural networks, deep iterations as well as fully supervised, weakly supervised, and unsupervised learning) to process-based statistics (such as methods in intelligent diagnosis, efficacy assessment, and prognostic applications) and detailed comparative studies based on organ, such as diseases of the brain, eye, breast, lung, bone, and skin.
3. Based on the above-mentioned in-depth analysis, we summarize the challenges faced by MMDLM applications and present directions for future scholars.
4. We analyze literature metrology, cited journals, and literature as well as keywords to arouse the interest of researchers in multi-modal deep-learning applications in the medical field. Furthermore, periodical distributions are discussed to effectively help scholars search for related research topics. Co-author, co-occurrence, co-citation, and literature coupling analyses are also performed.

The scope of the discussion in this review is shown in Fig. 1.

### 1.3 Structure of this Paper

The remainder of the paper is organized as follows: Sect. 1 describes the development of the multi-modal concept and some of the related problems encountered thus far. Section 2 reviews the progress of multi-modal deep learning in medical applications. Section 3 describes the multi-modal fusion pre-training algorithms, medical multi-modal fusion methods, their performances, and a comparative study of their key features. Section 4 describes the use of the VOSviewer software as well as the co-authorship, co-citation, co-occurrence, and literature coupling analyses that were performed in this study. Finally, the important conclusions thus derived are discussed in Sect. 5.

## 2 Literature Review

### 2.1 Multi-modal Medical Applications

A modality is a particular mode wherein something exists, is experienced, or is expressed. When a research problem comprises several modes, it is characterized as a multi-modal research problem. Simultaneously, modes can also be defined in a very broad manner. For example, data regarding two different languages, or datasets collected under two different circumstances, can be regarded as two modes [4]. To better understand the world around us, we should be able to interpret multiple signals simultaneously. For example, images are often associated with labels and text explanations, and text often contains images for the clear expression of the central idea of an article. Considering that different modes have different statistical properties, MMDLM can be used for processing and understanding multi-source modal information using deep-learning techniques.

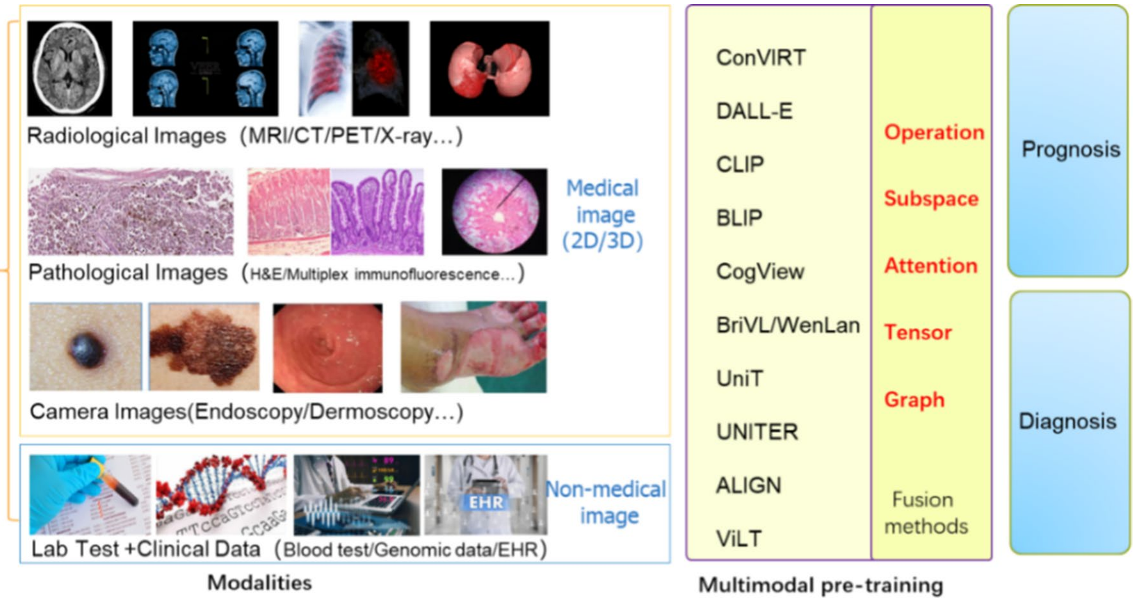


**Fig. 1** Scope of discussion in this review

At present, multi-modal learning for image, video, audio, and semantics is being deeply investigated. Particularly, research is being conducted on deep neural networks to learn multi-layer representations and for the abstraction of data before it is converted into high-level abstract features of the network. Image analysis has made important research progress in various medical fields such as classification, segmentation, detection, and localization [4]. Deep convolutional networks have been actively used in medical image analysis in areas such as segmentation, anomaly detection, disease classification, computer-assisted diagnosis, and retrieval. For a long time, medical imaging has been a diagnostic method in clinical applications. Recent advances in hardware design, security programs, computing resources, and data storage capabilities have significantly benefited the field of medical imaging [2]. Currently, the main application areas of medical image analysis include the segmentation, classification, and anomaly detection of images generated using a wide range of clinical imaging modes. For example, in 2021, Qian's team from the University of Southern California proposed a deep-learning system based on multi-modal and multi-angle medical ultrasound images, and successfully verified the accuracy, robustness, and effectiveness of the system in the prospective clinical environment of several hospitals. The results showed that the interpretable artificial intelligence (AI) assisted diagnosis system can significantly optimize the diagnosis results of human doctors, improve the clinical applicability of its auxiliary diagnosis, and provide new ideas for subsequent clinical translational research [8]. With the rapid development of medical information technology and modernization of medical equipment, an enormous amount and variety of medical data has emerged. Medical data can be broadly classified into three main categories according to the specific information and forms they present:

1. Clinical text data, which mainly includes structured test data such as hemoglobin and urine routine, as well as unstructured text data such as patient complaints and pathology texts recorded by doctors.
2. Image and waveform data, including imaging data such as ultrasound images, CT images, MRI images, and signal data such as ECG and EEG.
3. Biomics data, which can be subdivided into genomic, transcriptomic, proteomic, and other categories according to different molecular levels.

Each type of patient-related data is a data modality, and the different modalities of medical data provide information about the patient's diagnosis and treatment from a specific perspective. They contain overlapping and complementary information, further improving the accuracy of diagnosis and treatment by combining multiple types of medical information.

## 2.2 Multi-modal Pretraining

Multi-modal recognition is a method that extracts the complementarity between different modes, such as assisting physicians in diagnosis, the core of which lies in the fusion of medical images and texts (electronic medical records, laboratory reports, etc.). Multi-modal matching focuses on how to align two modal features, images, and texts. Table 1 shows the main studies, key application areas, and methods of common multi-modal pre-training for comparative analysis. To complete the medical multi-modal fusion method and performance comparison research, we conducted a

**Table 1** Mainstream multi-modal pre-training models

| | Study | Model | Keywords | Applications |
|---|---|---|---|---|
| 1 | Zhang et al. [37] | ConVIRT | Comparative learning, medical vision, unsupervised | Visual representation learning of medical images |
| 2 | Ramesh et al. [38] | DALL-E | Image generation transformer, codebook | text → Image + |
| 3 | Radford et al. [9] | CLIP | Contrast learning, feature space alignment | Text and text matching |
| 4 | Li et al. [39] | BLIP | Captioning and filtering (CapFilt) bootstrapping | Understanding and generation |
| 5 | Ding et al. [11] | CogView | Transfer-learning, Sandwich-LN, VQ-VAE, GPT | Chinese image generation |
| 6 | Huo et al. [40] | WenLan | Weak semantic correlation, Roberta, InfoNCE, MoCo, Faster-RCNN | Chinese text and text retrieval |
| 7 | Hu et al. [41] | UniT | Multi-modal, multitask, Transformer, DETR, BERT | Natural language understanding and multi-modal reasoning |
| 8 | Chen et al. [13] | UNITER | Integration, BERT, Faster R-CNN, multi-modal | Image and text matching |
| 9 | Li et al. [16] | ALIGN | Billion image-text pairs of noisy datasets, comparative learning | Align images and text pairs for visualization and verbal representation |
| 10 | Kim et al. [17] | ViLT | Image cutting, splicing, fast speed, the simplest multi-modal | Visualization, text |
| 11 | Yao et al. [42] | CPT | Alternatively, colorful prompt tuning | Pretraining vision-language models |

detailed comparison and survey of the literature. There are currently two main approaches to multi-modal tasks: light fusion and heavy fusion. The light fusion approach is usually effortless, such as the vector inner product, as represented by CLIP [9]and ALIGN [10], which use a two-tower structure focusing on multi-modal alignment to facilitate text matching, retrieval, and other downstream tasks. The heavy fusion approach is based on pre-trained Transformers [11], as represented by OSCAR [12], UNITER [13], VINVL [14], etc. These methods can be regarded as a single-tower structure that focuses on incorporating multi-modal information with an attention mechanism to perform additional tasks. Heavy fusion can interpret VQA [15], captions, and other downstream tasks that require information fusion and understanding, which the ALIGN algorithm [16] cannot perform. However, this approach is not as efficient as CLIP [9] in retrieval. Ultimately, the algorithm depends on the task. At present, there is a trend of unifying the two methods: the two-tower model, which is used as the base, and the single-tower model, which is incorporated in the upper layer. Alignment is performed before fusion. Multi-modal fusion refers to incorporating the information of multiple modes for classification or regression tasks. The benefits of multi-modal fusion are as follows: (1) more robust inference results can be generated for different modal representations of the same phenomenon, (2) auxiliary information that is not visible in a single-mode can be retrieved from multiple scales, and (3) for a multi-modal system, modal fusion can operate normally, even when a certain mode disappears. The modal and optimization methods used by the neural network for modal fusion may be different, however, the concept of information fusion through collaborative hidden layers is the same. Neural networks are also used for sequential multi-modal fusion and usually use RNNs (Recurrent Neural Networks) and LSTM (Long Short-Term Memory). Typical applications are audio-visual emotion classification, electronic medical records, etc. Some advantages of a deep neural network for modal fusion are as follows. It can (1) learn from large amounts of data, (2) perform end-to-end learning of multi-modal feature representation and fusion, and (3) performs better than non-deep-learning methods and can learn a complex decision boundary [4]. Table 1 shows that the multi-modal pre-trained model and its variants update and iterate very quickly, covering keywords including contrast learning, text and text matching, feature space alignment, understanding and generation, Chinese image generation, and transfer learning. These key technologies can be widely used to assist medical clinical applications, such as wearable device-based multi-source data health monitoring, human–machine automatic health assessment, and machine-based surgical safety assessment. To develop lightweight models and products that can adapt to complex medical data from multiple sources and assist clinical applications as soon as possible, research institutions (e.g., universities) and leading companies (e.g., Google) are conducting relevant research. Recently, multi-modal pre-trained models based on the Transformer structure have become very popular. Pre-training can be carried out using a large amount of unlabeled data, and then fine-tuning can be done with a small amount of labeled data. For example, as the simplest single-mode model, ViLT [17] is fast in visual and text processing, which provides a good foundation for embedded devices such as clinical consultations and surgeries. The approach used for multi-modal fusion depends on the task and data, and existing work often proposes various fusion methods without any real unified theoretical support. To efficiently and automatically select the fusion strategy according to the task or data, a neural architecture search (NAS) [18] is highly effective. Of course, there are challenges in multi-modal fusion; for example, the sequential information of different modes may not be aligned.

### 2.3 Medical Imaging and Non-imaging Models

Table 2 shows the research objects and tasks of medical image and non-image modalities. El-Sappagh et al.'s (2020) research has achieved excellent clinical results. They proposed a deep-learning fusion model based on Bidirectional Long Short-term Memory (BiLSTM) networks and Convolutional Neural Networks (CNNs). The multi-modal multitask model, based on five modalities [i.e., Magnetic Resonance Imaging (MRI), PET Positron Emission Computed Tomography, neuropsychological data, cognitive score data, and evaluation data], jointly predicts variables such as Alzheimer's disease (AD) multistage progression tasks and four key cognitive scores [19]. Table 2 also shows that The Cancer Genome Atlas (TCGA), The Alzheimer 's Disease Neuroimaging Initiative (ADNI), and other open-source datasets remain the first choice of most researchers. However, we can also observe that open-source data for medical multi-modal applications remain comparatively uncommon, which may be influenced by medical ethics. MMDLM is an important research direction for researchers in cancer prevention and therapy, which is related to the difficulty of cancer prevention and treatment [7, 20–25].

### 2.4 Deep Multi-modal Fusion Methods and Performance

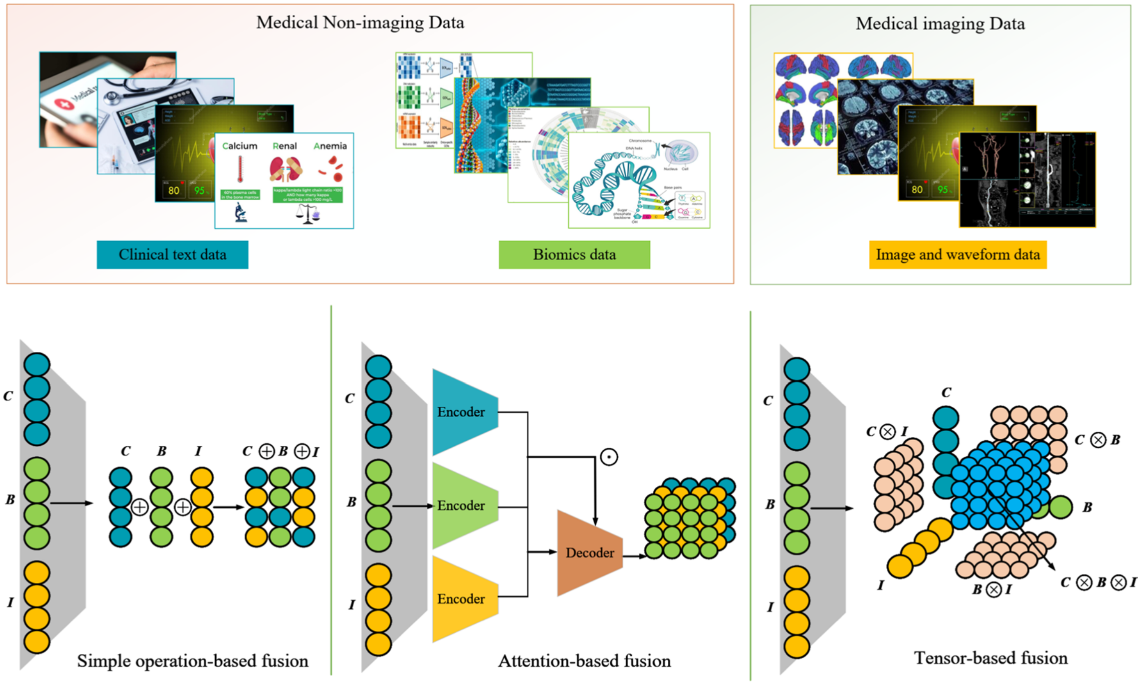
Multi-modal fusion is a key research point in multi-modal research. It integrates information extracted from different modes into a stable multi-modal representation. Multi-modal fusion is related to representations, and a process that focuses on using some architecture to merge representations of different single modes is classified as fusion. Fusion methods can be divided into late and early fusion according

**Table 2** Number of objects and tasks studied in medical imaging and non-imaging models

| | Study | Modalities | Subjects | Tasks |
|---|---|---|---|---|
| 1 | Zhang et al. [43] | CT | 654 contrast-enhanced CT/285 patients BRATS 2018 (Public) | Liver tumor and tumor segmentation |
| 2 | Fetit et al. [65] | Retinal, genomic, and clinical features | 3,891 individuals | Predicting the risk of diabetic cardiovascular and cerebrovascular disease |
| 3 | Holste et al. [20] | MRI, clinical features | 17,046 samples of 5,248 women | Classification of breast cancer |
| 4 | El-Sappagh et al. [19] | MRI, PET, neuropsychology data, cognitive scores, assessment data | 1,536 patients from ADNI (Public) | Classification of Alzheimer's disease |
| 5 | Yan et al. [21] | Pathological images, clinical features | 3,764 samples of 153 patients (Public) | Classification of breast cancer |
| 6 | Mobadersany et al. [27] | H&E, genomic data | 769 patients in the open datasets TCGA-GBM and TCGA-LGG | Survival prediction of patients with glioma tumors |
| 7 | Yap et al. [28] | Macroscopic images, dermatoscopic images clinical features | 2,917 samples | Classification of skin lesions |
| 8 | Silva et al. [32] | H&E, mRNA, miRNA, DNAm, Copy number variations, clinical features | 11,081 cases of 33 tumors in the open-source dataset TCGA | Pan-cancer survival prediction |
| 9 | Yoo et al. [30] | MRI, clinical features | 140 patients | Classification of brain lesion conversion |
| 10 | Yao et al. [22] | Pathological images, genomic data | 106 patients from TCGA-LUSC, and 126 cases from the open-source dataset TCGA-GBM | Survival prediction of patients with lung cancer and brain cancer |
| 11 | Cheerla et al. [23] | Pathological images, gene expression, microRNA expression, clinical features | 11,160 patients from the open-source dataset TCGA | Survival prediction of patients with 20 types of cancer |
| 12 | Li et al. [24] | Pathological images, genomic data | 826 cases from the open-source dataset TCGA-BRCA | Survival prediction of patients with breast cancer |
| 13 | Zhou et al. [44] | CT, clinical data, lab tests | 733 patients | Classification of COVID19 severity |
| 14 | Chauhan et al. [22] | X-rays, medical imaging reports | 6212 patients from the mimic-CXR open-source dataset | Three-level severity classification of pulmonary edema |
| 15 | Schulz et al. [25] | CT, genomic data, H&E, MRI | 230 patients from an open-source dataset and 18 patients from an external trial group | Survival prediction of patients with renal cell carcinoma |
| 16 | Cui et al. [8] | CT, clinical features | 397 patients | Evaluation and prediction of metastasis of lymphocytic carcinoma |
| 17 | Li et al. [45] | MRI, genomic data demography features | 112 patients | Efficacy prediction of chemotherapy in breast cancer patients |
| 18 | Guan et al. [31] | CT, clinical features | 553 patients | Classification of esophageal fistula risk |
| 19 | Wang et al. [7] | Pathological images, genomic data | 345 patients from TCGA (Public) | Survival prediction of patients with breast cancer |
| 20 | Zhou et al. [46] | PET, MRI, SNP | 805 cases from the open-source dataset ADNI | Classification of Alzheimer's disease and its prodromal status |

to their different locations. Since early and late fusion will inhibit inter-model interactions, current research focuses on intermediate fusion methods, which enable these fusion operations to be placed in multiple layers of the deep-learning model. There are three methods for the fusion of text and image: simple operation-based, attention-based, and tensor-based approaches. Figure 2 shows the schematic structure of the three multi-modal fusion methods.

Fusion is based on integrating feature vectors from different modes in simple ways, such as vector splicing, vector weighted sum, and so on. For multi-modal tasks, the first approach that comes to mind should be based on simple operations [19–21, 26–30], and the fusion of this approach has achieved the desired results in medical multi-modal applications. Holste et al. studied 10,185 breast enhancement MRI (DCE-MRI) data from 5248 women. They extracted clinical indications and breast density information from mammograms using a multi-modal model to reduce these data to 2D maximum intensity projections and then linked them to 18 non-image features, and they achieved an Area



*C*:**Clinical text data.** Mainly includes structured test data such as haemoglobin and urine routine, as well as unstructured text data such as patient complaints and pathology texts recorded by doctors.
*B*:**Biomics data.** This can be divided into genomic, transcriptomic, proteomic, etc., according to different molecular levels.
*I*:**Image and waveform data.** Including imaging data such as ultrasound images, CT images, MRI images and signal data such as ECG and EEG.

**Fig. 2** Schematic diagram of the structure of the three multi-modal fusion methods

Under Curve (AUC) of 0.849 [20]. This shows that simple operation-based methods with excellent datasets can also produce promising results. However, this method still has a problem; there is not enough interaction between the two modes, and the coupling relationship is insufficient.

The attention-based fusion approach, which is based on an attention mechanism, has been widely used in multimodal medical applications. To enable the model to pay attention to text, such as text in medical images and medical records, the mechanism gives different parts of the image feature vector different weights according to the characteristics of the image and text features. This enables the model to extract effective features from multi-source data and leverage the advantages of the multi-modal fusion [13, 25, 31, 32]. For multi-modal tasks, Silva et al. proposed an end-to-end multi-modal deep-learning generalized prognostic prediction model that predicted survival rates for all 33 cancers studied in a TCGA program. The model also used more input data modes than those of previous studies, including histopathological sectional images of clinical information and different genomic data [32]. The multi-modal fusion method-based tensors, also known as the bilinear pooling fusion method, are predominantly used to fuse visual feature vectors and text feature vectors to attain a joint representation space [33]. Through the stepwise decomposition of the weight tensor, an efficient multi-modal fusion model can be achieved. In recent years, the most important multi-modal fusion methods have been attention-based and bilinear pooling methods using attention-based fusion and tensor-based fusion [13, 34]. Clinical decision-making in oncology involves multi-modal data such as radiological scans, molecular analysis, histological sections, and clinical factors. Braman et al. used a deep orthogonal fusion (DOF) model to predict the overall survival of patients with glioma from different multi-modal data. The model learning will come from the multi-parameter information on MRI, the biopsy of the mode (such as images of DNA sequencing, and/or H&E (Hematoxylin–eosin Staining) slides), and clinical variable information combined into an integrated multi-modal risk model, at the same time introducing a multi-modal organization loss, through a complementary embedded model to improve performance. When the DOF model predicted the overall survival of glioma patients, the median C-index was $0.788 \pm 0.067$, which is significantly better than the best-performing single-peak model (median C-index: $0.718 \pm 0.064$; $P = 0.023$) [34]. Faisal Mahmood's team used multi-modal deep learning to integrate and analyze whole-section images and genetic profiling data from 14 cancers. The algorithm predicts good and poor prognostic outcomes in different forms, predicts patient prognosis based on morphological and molecular levels at the disease and patient levels, combines spatial distribution of tumor, stromal and immune cells in the tumor microenvironment to synthesize and consider an open database has been established for further exploration, as well as for biomarker discovery and characterization [35]. Figure 3 displays a performance comparison of several researchers who conducted multi-modal deep learning using three different fusion modes.

Through an overall comparative analysis, the three-stage deep feature learning and fusion diagnosis framework proposed by Zhou et al. [46] is considered a good approach among the sample of studies we collected. The framework is mainly designed to identify Alzheimer's disease (AD) and its precursor states, and it progressively integrates multimodal imaging and genetic data at each stage, effectively alleviating the problem of multi-modal data heterogeneity. The framework also partially solves the problem of incomplete multi-modal data by designing a staged deep-learning strategy. Overall the framework achieves an all-round balance between the data and the fusion process. Based on our analysis, we believe that a balanced approach to data and fusion process is crucial for the success of multi-modal deep learning in medicine. Therefore, we recommend that future studies focus on gradually integrating or extracting the features of different modalities in a specific order during the fusion process, while also prioritizing data quality and progressively making it more complete.

In multi-modal deep learning, the process of collecting effective features from different modes is called "multimodal fusion". In this process, the several modes are not simply and separately given as input to the model. The fusion of the different data modes can occur at different stages of the process. For example, the simplest early fusion technique involves concatenating the input modes or features before the processing stage; however, this technique cannot be applied to complex data modes. A more sophisticated approach is intermediate fusion, wherein the representations of the different modes are combined and co-learned during training. It allows for modal-specific preprocessing while capturing interactions between data modes to achieve joint representations. Late fusion is also a simple method wherein a separate model is trained for each mode and combined with the output probability for joint representation. However, such a fusion method misses the opportunity to extract information from the interaction between the modes. Over the past few years, deep learning has transitioned from mode-specific architectures—such as the CNN for graphics or RNN for text—to the Transformer. This novel architecture performs well across various input and output modes and tasks. One promising aspect of the Transformer is its ability to learn meaningful representations from unlabeled data, as the resources required to obtain high-quality markup in the medical field are limited and expensive. At the same time, because of the restrictions associated with privacy protection, medical ethics, and other relevant rules, it is highly difficult to obtain medical data. One possible solution is to use the available data from one mode to aid learning using

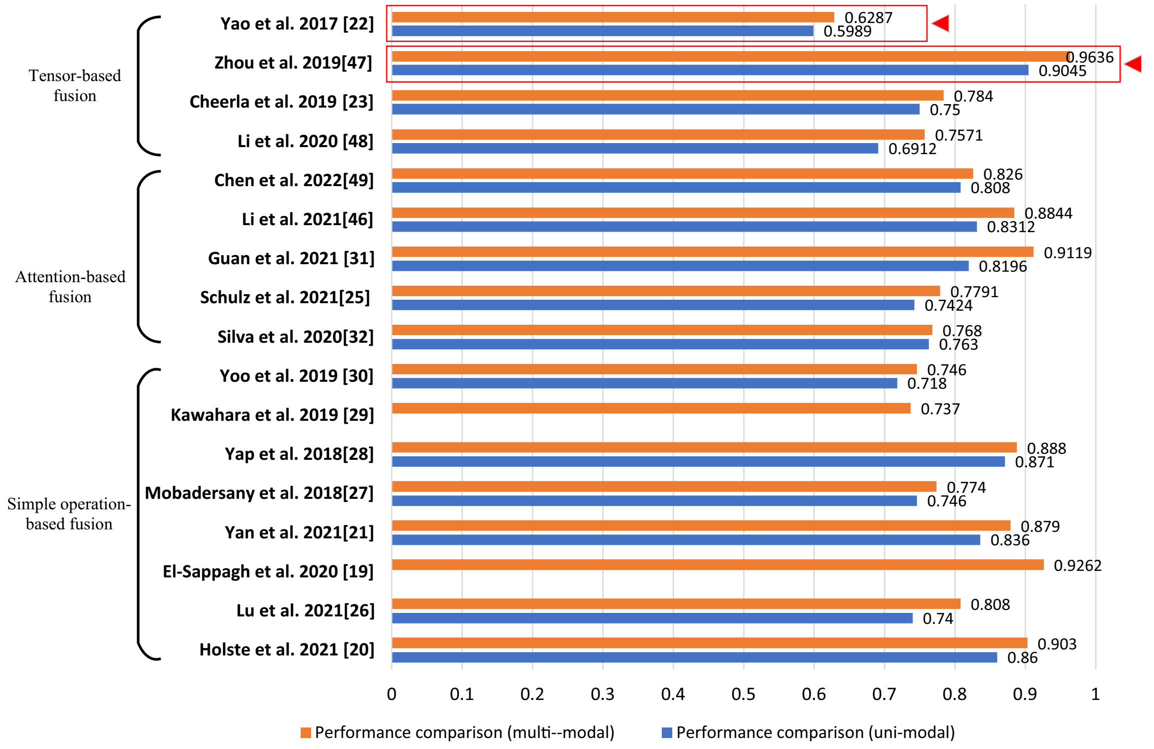


**Fig. 3** Comparison of multi-modal fusion performance (AUC/C-index/Accuracy)

another via a multi-modal learning task called "co-learning." For example, some studies have suggested that a Transformer that is pre-trained on untagged language data may generalize well to other tasks. In medicine, a model architecture called "CycleGans", which was trained using unpaired uncontrasted or contrast computed tomography (CT) scan images, is used to generate uncontrasted or contrast CT scan images [36].

## 3 Discussion

In the previous section, we reviewed the current models used for researching disease prognosis and diagnosis using deep learning-based approaches that merge images and non-images. Through a literature analysis, it was found that as of 2021, Transformers have assisted in the rapid emergence of multitask and multi-mode AI. OpenAI built the first model, called DALL·E [38], with a Transformer architecture that can process both image and text data simultaneously, known as the image version of GPT-3[49]. The CLIP model, which can pair text and images, was likewise introduced [9]. Facebook released a series of improved Transformer models, one of which is UniT [41], which can simultaneously handle two modes of data and seven tasks, such as natural language processing, natural language understanding, image recognition, and object detection. While most of these multi-tasking, multi-modal AI systems are in the research and experimental stage, some have already achieved good results in practical applications. For instance, electronic health records (EHR) have a complex multi-modal structure. Xu et al. used Neural Architecture Search (NAS) and Multi-modal Fusion Architecture Search (MUFASA) to select both single-mode and cross-mode network architectures. This approach outperformed the single-mode NAS on publicly accessible EHR datasets [37]. Feature-level fusion methods are further divided into operation-, tensor-, attention-, subspace-, and graph-based methods [17]. While the operation-based approach is intuitive and effective, it may lead to poor performance when learning complex interactions between different modes. The tensor-based approach has the risk of overfitting. The attention mechanism-based method is a quite effective method for multi-modal feature fusion and can calculate the inter-modal and intra-modal importance features; hence, it is widely used in multi-modal fusion applications. Furthermore, with the help of self-attention, the steps of modal fusion do not need to be designed carefully. One can simply splice the multi-modal information into a sequence and use a Transformer encoder to learn their binary relations and merge the information within and between modes simultaneously or run multiple modes in parallel. Then, the Transformer decoder can be used to perform the cross-mode fusion. The Transformer can easily handle input types with graphs, which are a more universal input structure. Both images and sequences can be converted into graphs, and hence, the Transformer is a more universal network structure. For example, Stanford University has created a set of open-source Transformer models called ConVIRT [37], which can automatically annotate X-rays with text. The self-attention mechanism does not require the user to explicitly designate the prior adjacency matrix, but simply input sufficient data (if available) and let the model learn the edge weights on its own. Compared with CNNs and RNNs, the Transformer has a larger number of parameters, stronger expression ability, and requires more training data. Accordingly, to effectively assist physicians in diagnosis and therapy, scientists should conduct comparative studies on multi-modal learning in the medical field and develop and share more benchmark datasets. Simultaneously, although multi-modal learning is advantageous for model performance optimization, research on modal selection should focus on calculating the model capacity, data quality, and specific tasks.

The successful advancement of multi-modal deep learning in medicine requires a large amount of data, and challenges are encountered when applying data, models, and performing complex tasks in this field. We list the main challenges below.

1. The diversity and uncertainty in medical datasets, including image and non-image data, sample size, depth of phenotypic analysis, heterogeneity and diversity of participants, degree of data standardization and harmonization, and degree of correlation among data sources together constitute a greater challenge than that posed by a single-mode deep-learning model. The challenge of handling the highly variable data found in real-world clinical databases must therefore be considered to ensure effective application.
2. In the development of multi-modal medical research and clinical applications, the collection, linking, and cost-effective annotation of multi-dimensional medical data also leads to challenges in terms of cost and speed.
3. In multi-modal medical fusion, it is necessary to properly associate all the data types in the dataset and extract effective feature sets. However, small and incomplete datasets as well as non-standardized data structures, which are prevalent in this field, pose significant challenges.
4. For some modes (e.g., 3D imaging and genomics), processing even a single point in time or individual instance of data requires a large amount of computing power. Hence, building models to simultaneously and rapidly process large-scale tumor pathological slides, genomics, or medical text data is an important fundamental challenge.

5. When collecting health and clinical data for research, privacy concerns can be raised by patients and doctors. Establishing a trusted mechanism to monitor and mitigate these issues is critical, and it requires researchers to propose and explore more solutions.
6. Multi-modal medical data fusion analysis is a multidisciplinary field that requires repeated interactions among clinicians, statistical analysis engineers, algorithm engineers, bioinformatics engineers, and professionals from other disciplines to determine research schemes. This interaction is hindered by the significant challenges of collaboration.

To meaningfully process and integrate the information in different medical data and increase the participation of AI in the assisted diagnosis and treatment process, joint efforts between the medical community and AI researchers will be required to construct and validate new models and ultimately demonstrate their ability to improve diagnosis and treatment.

## 4 Bibliometrics

Bibliometrics is an effective quantitative method for examining research activities in a specific field. To describe, evaluate, and monitor MMDLM-related research results, we conducted a comprehensive analysis based on annual trends, countries, institutions, journals, highly cited papers, co-citations, coauthors, co-occurrence analyses, and literature coupling [17].

### 4.1 Literature Resources

To investigate this topic, we selected literature that adheres to the principles of representativeness and universality. The Web of Science (WoS) Core Collection database was employed as the data source for the retrieval of the literature related to multi-modal fusion and deep learning based on medical images. For the query, we used TS = (Multi-modal deep learning (OR (Multi-modal deep learning)) AND medicine). A total of 920 records were retrieved on this subject. Simultaneously, we used the refining function of the WoS to extract the key information of the remaining publications, eliminate irrelevant or weakly related publications, and select a final total of 879 publications as the basic target data of this study. The literature was collected from January 2010 to April 2022. All the selected publications met the following criteria: (1) the research content (in part/in whole) presented a multi-modal deep-learning fusion approach for medical applications and (2) the publication focused on the improvement of the deep-learning algorithm and its application in one or more detailed medical fields.

### 4.2 Analytical Methods

Considering the large number of publications identified in this study, it would have been challenging to manually extract the information individually and further explore the relationship between them. Therefore, it was necessary to sort them using a bibliometric analysis, which aims to study the distribution structure, quantitative relationships, and variation in the literature using measurement methods such as mathematics and statistics [50]. Common bibliometric mapping software tools include HistCite [51], CiteSpace [52], and VOSviewer [53, 54]. By comparing the features of these software tools, we found that VOSviewer is easier to use than the other software. By constructing the relationship and visually analyzing of network data (mainly literature knowledge units), the software can render a scientific knowledge graph and show the structure, evolution, cooperation, and other relationships in the knowledge domain. Its outstanding feature is its strong graphical display ability, which is suitable for large-scale data. We used VOSviewer to explore the collaborative publications and networks of multi-modal fusion in medical image research, as well as the collaboration and distribution among countries, research institutions and authors on the subject. We set the node types to “country”, “agency”, and “author” in VOSviewer software. At the same time, considering the close relationship between countries and institutions, we added nodes for countries and institutions in the same graph. By setting the node type to “category” and selecting the timeline view of the software, we visualized the evolution of the objects studied in this domain. When the node type is set to “Reference”, co-citation analysis can be carried out to analyze representative research articles. In addition, when the node type is set to “Terms”, cluster analysis is conducted according to the nodes in the co-occurrence graph. Ultimately, cluster analysis and term co-occurrence are used to distinguish the frontiers in research and hotspots in the different development stages of this research field.

### 4.3 Literature Analysis

#### 4.3.1 Trends in the Number of Published Papers

We investigated 879 papers that were published between 2010 and 2022 (Fig. 4). The growth in the number of publications can be divided into three stages, which we call preparation, rise, and prosperity.

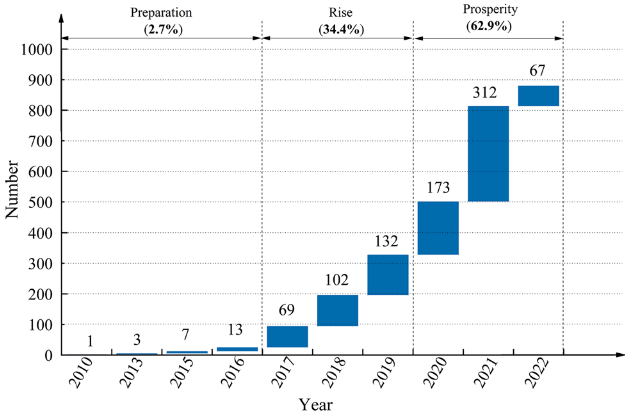


**Fig. 4** Changes in the number of MMDLM publications from 2010 to 2022

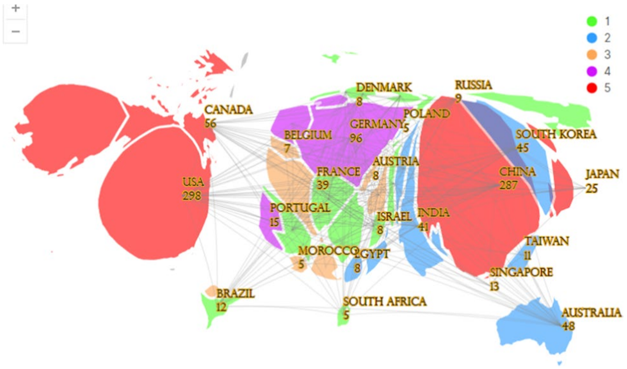


**Fig. 5** Author distribution by country

#### 4.3.2 Annual Trends and Possible Explanations

By analyzing the collaboration networks between different countries and institutions, we can clearly see the achievements of countries and research institutions in the field of multi-modal fusion skills, and further identify the collaboration between them by analyzing relevant publications. Figure 4 shows the annual change in the number of multi-modal deep-learning medical publications.

With the widespread application of artificial intelligence in medicine, the number of MMDLM publications continues to expand. It saw particularly strong growth in 2017. There are two possible reasons for this rise:

1. Many researchers have begun to use deep learning to solve medical problems.
2. The pioneering research achievements globally have piqued the interest and increased the confidence of researchers in this approach.

#### 4.3.3 Countries

The analysis of literature data shows that scholars from 35 countries/regions have published publications about MMDLM, but over 85% of the publications were contributed by scholars from five active countries, as shown in Fig. 5. The United States was the largest contributor to MMDLM publications, with its scholars contributing 298 publications or 33.9%. The contribution of Chinese scholars to publications in this field ranks second in the world, accounting for 32.65%. After the United States and China, Germany, the United Kingdom, and Canada ranked third, fourth, and fifth with 10.92%, 9.88%, and 6.37%, respectively.

#### 4.3.4 Institutions

After calculating the outstanding publication contributors at the national level, we further analyzed the outstanding publication contributors at the institutional level. According to the collected data, the League of European Research Universities (LERU) is a prominent institution that has contributed 79 publications, accounting for 8.957% of the global publication volume on this topic. As an important contributing country, China ranked second and third in terms of the intellectual output of the University of Chinese Academy of Sciences and Shanghai Jiao Tong University, with 36 and 33 publications, respectively. Table 3 shows the details of the top-16 most productive organizations that actively generated MMDLM-related publications. Of the top-16 institutions, the main countries/regions with which these institutions are affiliated are the United States (6), China (4), the United Kingdom (2), LERU (1), Germany (1), and France (1).

#### 4.3.5 Highly Cited Papers

To identify the most influential development ideas and scientific thinking in MMDLM research, we selected 12 highly cited papers (HCPs) and 1 "hot" paper from the WoS and ranked them according to their total citation frequency. Table 4 lists authors, journal names, regions, titles, and citations for these HCPs. The MMDLM-related HCP proposed by Arbabsiar [55], which has been cited 381 times, can be regarded as a pioneering work in the single-discipline prediction of brain dysfunction based on neuroimaging. Emerging trends such as multi-modal brain imaging, survival prediction, disease subtype classification, and multi-modal attention mechanisms are also discussed. Emerging imaging digitalization technologies and data-intensive computational methods such as reinforcement learning meet the needs of post-operative brain tumor prediction and evaluation applications. The paper, published in *NeuroImage* (a top journal), has been cited more than any other paper in the analysis. HCPs have proposed different types of multi-modal

**Table 3** Author distribution by institution

| Rank | Institution | Regions | Publications | Share (%) |
|---|---|---|---|---|
| 1 | League of European Research Universities | LERU | 79 | 8.957 |
| 2 | Chinese Academy of Sciences | China | 36 | 4.082 |
| 3 | Shanghai Jiao Tong University | China | 33 | 3.741 |
| 4 | Harvard University | USA | 30 | 3.401 |
| 5 | University of London | UK | 30 | 3.401 |
| 6 | Technical University of Munich | Germany | 24 | 2.721 |
| 7 | University of California System | USA | 23 | 2.608 |
| 8 | North Carolina State University | USA | 23 | 2.608 |
| 9 | University of Texas System | USA | 23 | 2.608 |
| 10 | Zhejiang University | China | 23 | 2.608 |
| 11 | Harvard Medical School | USA | 22 | 2.494 |
| 12 | Fudan University | China | 21 | 2.381 |
| 13 | The University of North Carolina at Chapel Hill | USA | 20 | 2.268 |
| 14 | Imperial College London | UK | 19 | 2.154 |
| 15 | Université Fédérale Toulouse Midi-Pyrénées | France | 19 | 2.154 |

*LERU* League of European Research Universities, including 23 universities in Europe

The University of California System is a consortium of 10 universities located in California

**Table 4** Highly cited papers

| Rank | Authors | Region | Paper title | Abbreviated journal title | Citations |
|---|---|---|---|---|---|
| 1 | Arbabshirani et al. [55] | USA | Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls | *NeuroImage* | 381 |
| 2 | Ibtehaz et al. [6] | Bangladesh | MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation | *Neural Networks* | 335[a] |
| 3 | Araujo et al. [24] | Portugal | Classification of breast cancer histology images using convolutional neural networks | *PLOS ONE* | 320 |
| 4 | Shin et al. [56] | UK | Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data | *IEEE T. Pattern. Anal* | 327 |
| 5 | Zhao et al. [57] | China | A deep learning model integrating FCNNs and CRFs for brain tumor segmentation | *Med. Image Anal* | 316 |
| 6 | Mahmud et al. [58] | Italy | Applications of deep learning and reinforcement Learning to Biological data | *IEEE T. Neur. Net. Learn* | 282 |
| 7 | Estai et al. [59] | Australia | Best teaching practices in anatomy education: A critical review | *Ann. Anat* | 262 |
| 8 | Liu et al. [60] | Australia | Multi-modal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease | *IEEE T. Bio-Med. Eng* | 258 |
| 9 | Bychkov et al. [63] | Finland | Deep learning-based tissue analysis predicts outcome in colorectal cancer | *Sci. Rep* | 235 |
| 10 | Arevalo et al. [61] | Colombia | Representation learning for mammography mass lesion classification with convolutional neural networks | *Comput. Meth. Prog. Bio* | 214 |
| 11 | Liu et al. [62] | China | The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges | *Theranostics* | 219 |
| 12 | Zeng et al. [64] | USA | DeepDR: A network-based deep-learning approach to in silico drug repositioning | *Bioinform* | 175 |

Highly cited papers are the top 1% of cited papers in the past decade

[a] Hot paper: a paper published on this subject in the past two years and that received enough citations to put it in the top 0.1% of academic papers

deep-learning methods for solving medical processes, such as different fusion models for physical medical image segmentation [6], brain tumor segmentation [24], multi-organ detection [56], anatomical education [57], multiple diagnoses of Alzheimer's disease [58], the prognosis of rectal cancer [59], breast mass detection [60], accurate diagnosis

[61], and the treatment of tumors [62]. These theories and techniques are considered an indispensable part of MMDLM research.

### 4.3.6 Research Landscape

MMDLM research is not restricted to the “Medical” and “Computer Science” domains but covers 93 WoS categories. This manifests in the widespread application of MMDLM theories and approaches in various domains. “Radiology Nuclear Medicine Medical Imaging”, “Engineering Biomedical”, “Imaging Science Photographic Technology”, and “Artificial Intelligence” are the biggest categories, containing approximately 50% of the related documents. Figure 6 shows the primary WoS categories that belong to MMDLM-related documents. In addition to the “Computer Science” and “Medical” categories, MMDLM-related documents were also seen in the “Neurosciences”, “Mathematical Computational Biology”, “Multidisciplinary Sciences”, “Optics”, “Telecommunications”, “Clinical Neurology”, “Neuroimaging”, “Oncology”, “Instruments Instrumentation”, and “Biochemical Research Methods”, categories. This is a profound demonstration of the wide application of MMDLM.

### 4.3.7 Keyword Co-occurrence Analysis for MMDLM

Keyword co-occurrences can effectively reveal research hotspots in this field. To explore the research hotspots of MMDLM, VOSviewer literature analysis software was used to perform bibliometric analysis on the keyword co-occurrences of the 879 analyzed studies. We obtained a total of 186 keywords from MMDLM-related publications. VOSviewer visualization software was used to simulate the MMDLM keyword co-occurrence network (Fig. 5). Each node in the visual platform keyword representing density has a special color, which is closely related to the link density on the node, and the color of the node depends on the degree of closeness of the node neighbor relationship. The red component in the keywords indicates its frequency is high. In contrast, keywords that appear less frequently have an amber color. Density visualization is very effective for understanding the overall structure and focusing on the most important components. Table 5 shows the first 13 keywords of MMDLM-related publications. According to Fig. 7, we

**Table 5** Top-10 keywords of MMDLM-related publications

| Rank | Keyword | Total link strength | Occurrences |
|---|---|---|---|
| 1 | Deep learning | 1395 | 451 |
| 2 | Classification | 518 | 119 |
| 3 | Segmentation | 312 | 86 |
| 4 | MRI | 275 | 63 |
| 5 | Diagnosis | 210 | 51 |
| 6 | Prediction | 206 | 39 |
| 7 | Magnetic resonance imaging | 197 | 34 |
| 8 | Feature extraction | 180 | 32 |
| 9 | Model | 145 | 34 |
| 10 | Images | 141 | 33 |
| 11 | Cancer | 129 | 32 |
| 12 | Alzheimer’s disease | 115 | 21 |
| 13 | Features | 114 | 25 |

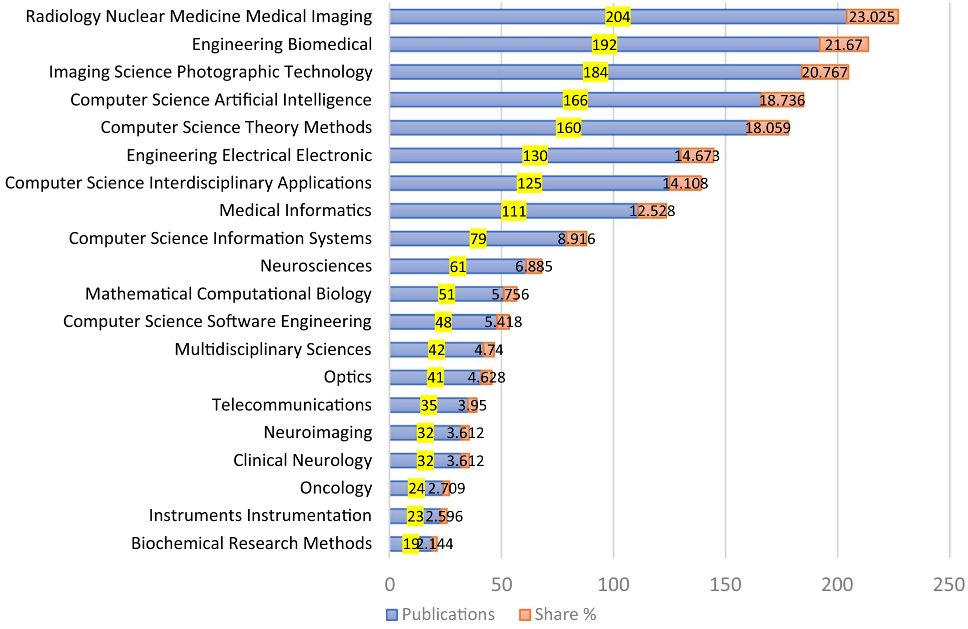


**Fig. 6** MMDLM research categories

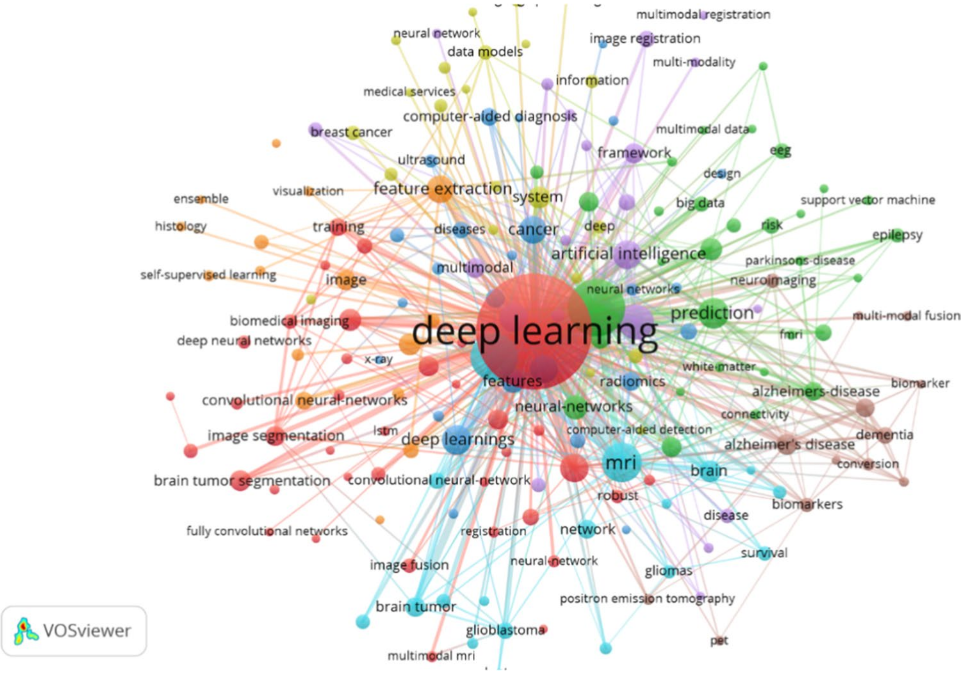


**Fig. 7** Keyword co-occurrence network of MMDLM-related publications

can intuitively identify the following research hotspots of MMDLM:

1. Deep learning plays a core role in medical multi-modal applications, which should be related to computer science and technology disciplines in most of the literature.
2. At present, researchers mainly focus on the classification and segmentation of physiological data, which is a traditional deep-learning application.
3. Tumor and brain science are important areas of concern for researchers.

The node and phrase font sizes in Fig. 7 represent weights. As can be seen from the network diagram, the larger the node (keyword) font, the larger the corresponding weight is. At the same time, the Euclidean distance of two nodes indicates the strength of the coupling. A direct link between two keywords indicates that they occur simultaneously. The more densely connected the lines are, the more frequently they occur. The document analysis software VOSviewer splits the keywords of MMDLM-related publications into eight clusters according to the coupling relationship, and each cluster is represented by the same color. The keywords "deep learning" and "classification" appeared most frequently. Keywords with high frequency include "segmentation" (86), "MRI" (63), and "magnetic resonance imaging" (34). To represent the working frequency of two keywords, the link strength between two nodes represents the co-occurrence frequency strength, which shows the quantitative parameters of the coupling relationship between two nodes. Calculating the total link strength of a node is the sum of the link strength of that node and the link strength of all other associated nodes. By observing this intensity, it is possible to intuitively calculate the degree of closeness of correlated studies, which may lead researchers to pay attention to the subcategories of the topic research and determine future directions.

### 4.3.8 Co-citation Analysis for MMDLM-Related Publications

When two (or more) articles are simultaneously cited by one or more later articles, the two articles are said to have a co-citation relationship. In co-citation analysis, more representative literature is frequently selected as the analysis objects, and hence a network analysis method was adopted to perform cluster analysis on this literature. The knowledge graph of a research area can be intuitively calculated. At the same time, co-citation analysis is widely used to disclose the coupling of authors, literature, and journals in the research field. In this section, we introduce the co-citation of authors, literature, and journals in the medical applications of multi-modal deep learning. Figure 8 shows the journal collaboration network for MMDLM-related publications. By analyzing and summarizing the co-citation relationships between authors and publications in the research field, the author's citation network can be displayed, which can reveal the author's research interests. The VOSviewer software was used to draw the author atlas of MMDLM scholars, as shown in Fig. 9. Unsurprisingly, the node coupling degree indicates that the nodes of Km, Kamnitsas, and Bengio are the largest. It also shows that they actively participate in this field.

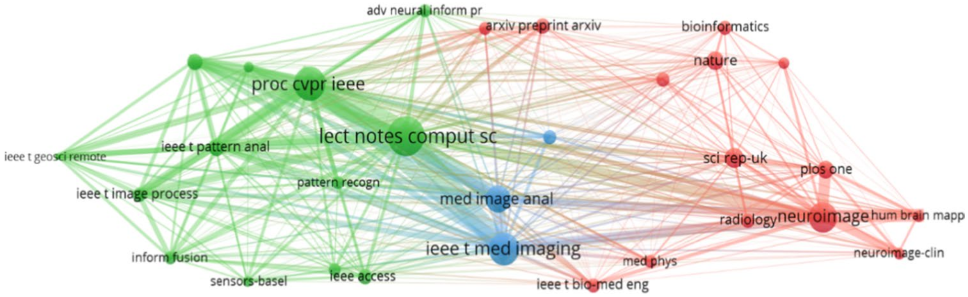


**Fig. 8** Journal co-authorship network of MMDLM-related publications

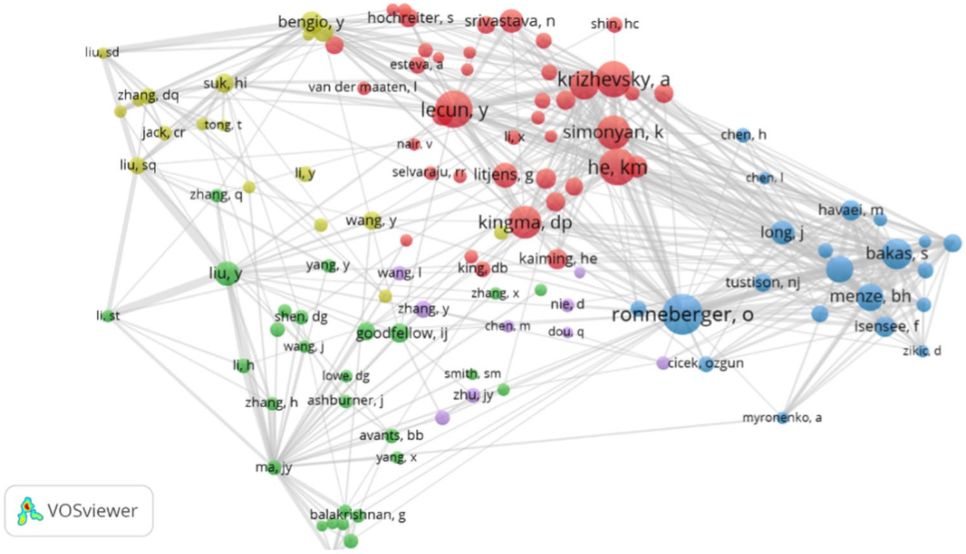


**Fig. 9** Co-authorship network of MMDLM-related publications

### 4.3.9 Bibliographic Coupling Analysis

In citation network, the bibliographic coupling of two articles refers to the number of the same articles in the references of both articles, that is, the number of other articles cited by both articles at the same time. We can define the undirected document coupling network corresponding to the directed citation network as follows: if two articles have at least one identical reference, there will be an edge between the corresponding two nodes. Literature coupling reflects the relationship between two cited references, while co-citation reflects the relationship between two cited references. Publication coupling analysis can show which domain the publications are more concerned with, and the weight correlation in coupling analysis depends on the number of reference publications they share. Consistent with the above coauthor analysis and co-citation analysis, we focus on discussing and showing the publication coupling network relationships of 879 articles by author (see Fig. 10), journal (see Fig. 11), institute (see Fig. 12), and country (see Fig. 13).

## 5 Challenges and Perspectives

The successful advancement of multi-modal deep learning in medicine requires a large amount of data, and challenges are encountered when applying data, models, and performing complex tasks in this field. We list the main challenges below.

1. Lack of standardized data collection and annotation protocols, which can lead to bias and limit the generalizability of models.
2. Interpretability of multi-modal deep-learning models is still a challenge, which may hinder their application in clinical practice.
3. Heterogeneity of data from different modalities, which may require different preprocessing and integration techniques to be effectively combined.
4. Availability and accessibility of multi-modal data, as it may require collaboration between different institutions

**Fig. 10** Author bibliographic coupling network of MMDLM-related publications

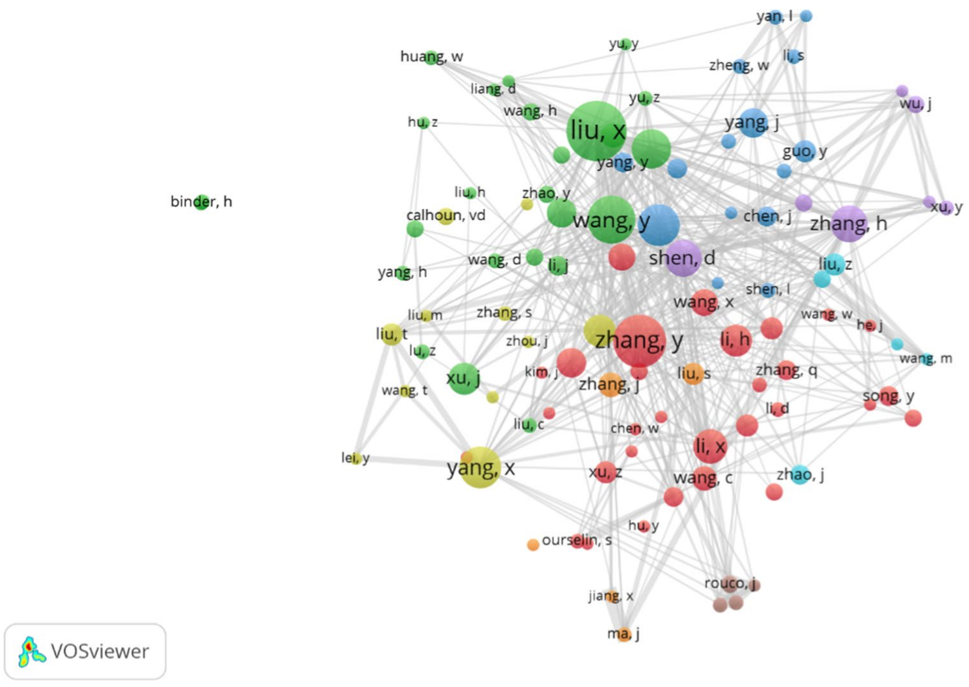
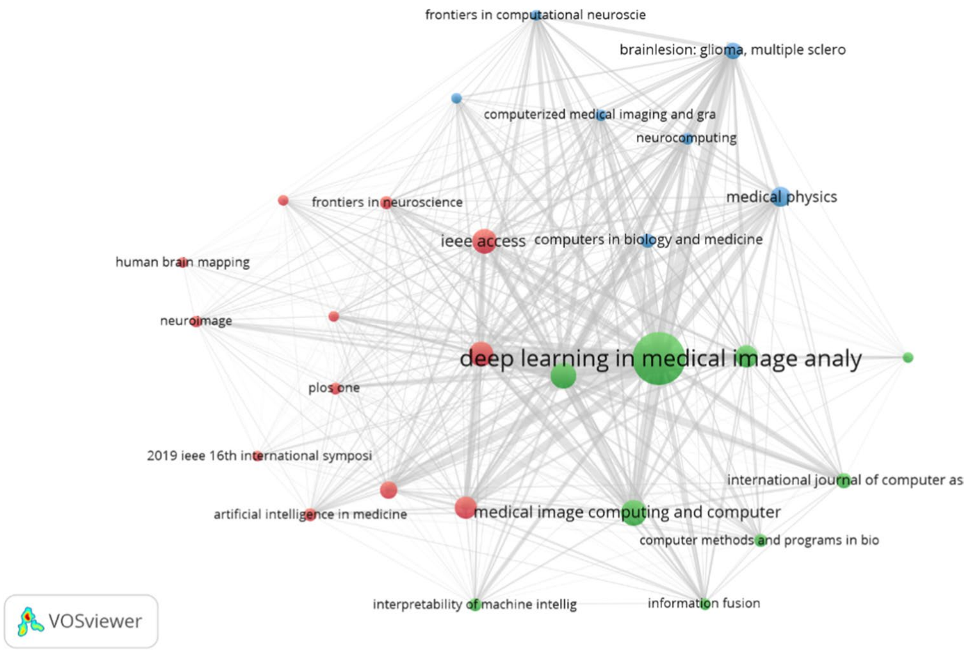


**Fig. 11** Journal bibliographic coupling network of MMDLM-related publications


frontiers in computational neuroscie
brainlesion: glioma, multiple sclero
computerized medical imaging and gra
neurocomputing
medical physics
frontiers in neuroscience
ieee access
computers in biology and medicine
human brain mapping
neuroimage
deep learning in medical image analy
plos one
2019 ieee 16th international symposi
international journal of computer as
artificial intelligence in medicine
medical image computing and computer
computer methods and programs in bio
interpretability of machine intellig
information fusion
VOSviewer

and data sharing agreements to collect and integrate data from multiple sources.

5. Model overfitting, which can lead to poor generalization performance on new data and be particularly problematic in medical applications.
6. Ethical implications of using deep-learning models in medical decision-making, such as potential biases and errors in the models, as well as the impact on patient privacy and autonomy.

A few visions

1. Multi-modal deep learning holds great promise for improving medical diagnosis and treatment, with the ability to generate more accurate and personalized predictions.
2. Integration of multi-modal data can improve the accuracy of diagnosis and prognosis of various diseases, such

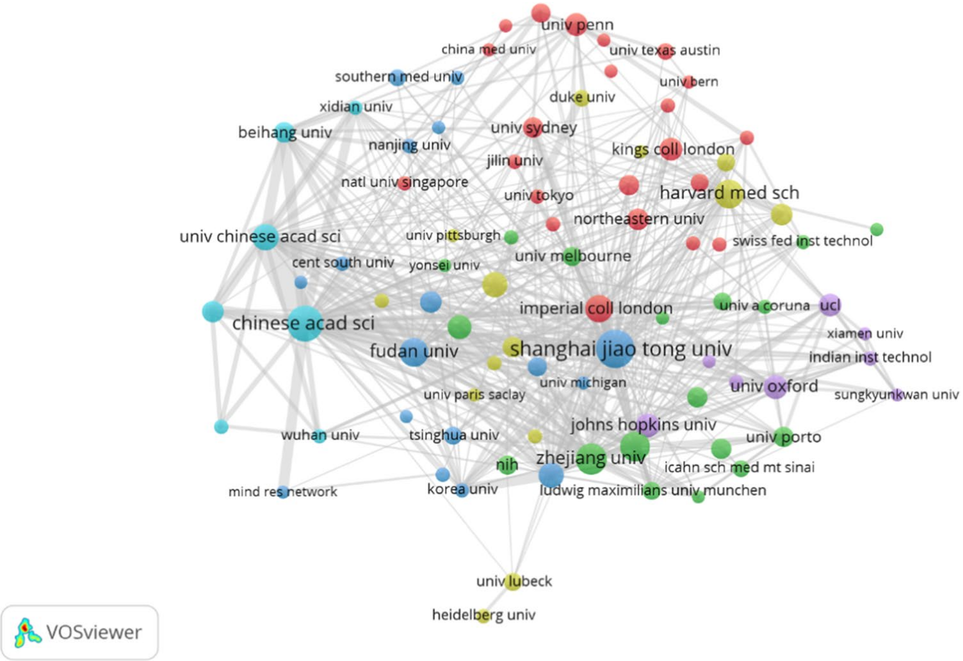


**Fig. 12** Institute bibliographic coupling network of MMDLM-related publications

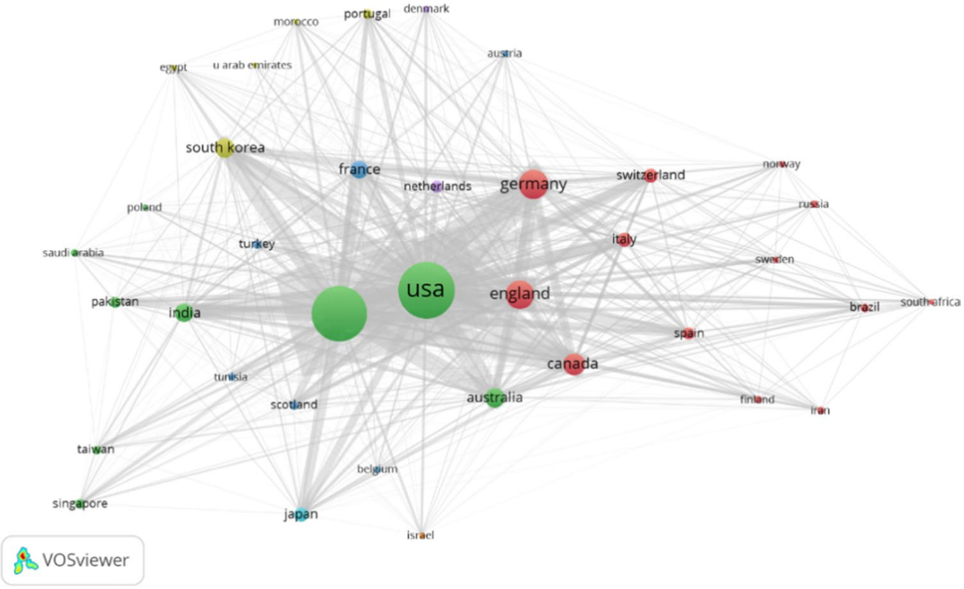


**Fig. 13** Country bibliographic coupling network of MMDLM-related publications

as cancer and Alzheimer's disease, and help clinicians predict disease risk and personalize treatment plans.

3. Multi-modal deep learning can provide more accurate and timely diagnoses and treatment recommendations, aiding clinical decision-making.
4. Approaches being explored include developing more standardized protocols for data collection and annotation, developing more interpretable and transparent models, and facilitating the sharing and integration of multi-modal data.

In summary, the application of multi-modal deep learning in medical research and clinical practice holds great promise for improving healthcare outcomes and personalized treatment. However, it is important to address the challenges associated with these models, including issues such as data heterogeneity, interpretability, and ethical considerations. By working together to address these challenges, we can unlock the full potential of multi-modal deep learning in healthcare, and improve the diagnosis, treatment, and outcomes for patients.

## 6 Conclusion

In this report, we comprehensively discussed the performance of medical multi-modal deep learning from the aspects of pre-trained networks, fusion approaches, and models in clinical and application studies, and we compared their key characteristics. In addition, we set out six significant challenges for multi-modal healthcare convergence. In the future, to obtain a clearer understanding of this new research trend, we will measure and evaluate the influential researchers, institutions, and research directions using bibliometric methods. From the results obtained in this study, we can conclude that the research can be approximately divided into three periods: preparation, from 2010 to 2016, rise, from 2017 to 2019, and prosperity, from 2020 to the present. Aiming at the current situation of insufficient polymorphic data fusion in deep learning in medical applications and to address the needs of human–machine collaborative medical cross-modal auxiliary diagnosis and treatment applications, we reviewed the use of deep learning to perform multi-modal medical data and medical knowledge fusion analysis research and establish a medical heterogeneous multi-dimensional data retrieval and matching mechanism. To this end, network methods were reviewed to provide improved multi-layer semantic feature matching of medical heterologous data, breakthrough multiple modal heterogeneous data fusion mechanisms, and perform multi-modal depth learning to provide a research and application basis to assist the doctors in their auxiliary diagnosis and treatment. Through the analysis of the collected research literature, we found that researchers are particularly concerned about cancer research based on artificial intelligence technology. In terms of scientific adaptation, they focus on diverse approaches to model integration. In addition, it is hoped that deep-learning models can be applied to long-term health monitoring and disease prevention to ensure sufficient data support for the in-depth development of AI in the medical field and further improve multi-modal systems. We believe that the findings of this report will play an important role in guiding and developing the importance of AI in clinical medicine and research.

**Acknowledgements** This work was funded by a scientific research project of the Science and Technology Department of Shanxi Province (Grant number 2020FP-11).

**Author Contributions** All authors contributed to the study's conception and design. Material preparation was performed by PXD. The first draft of the manuscript was written by PXD. and then polished by PBZ, ZK, and LY. All authors read and approved the final manuscript.

**Availability of Data and Materials** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Code Availability** Not applicable.

## Declarations

**Conflict of interest** We declare that we have no conflicts of interest.

**Ethics approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References

1. Du, J., et al.: An overview of multi-modal medical image fusion. Neurocomputing **215**, 3–20 (2016)
2. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
3. Atrey, P.K., et al.: Multi-modal fusion for multimedia analysis: a survey. Multimed. Syst. **16**(6), 345–379 (2010)
4. Baltrušaitis, T., et al.: Multi-modal machine learning: a survey and taxonomy. IEEE Trans. Pattern. Anal. Mach. Intell. **41**(2), 423–443 (2018)
5. Ramachandram, D., et al.: Deep multi-modal learning: a survey on recent advances and trends. IEEE Signal Process. **34**(6), 96–108 (2017)
6. Ibtehaz, N., et al.: MultiResUNet: rethinking the U-Net architecture for multi-modal biomedical image segmentation. Neural Netw. **121**, 74–87 (2020)
7. Wang, Z., et al.: GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. Bioinform. **37**(18), 2963–2970 (2021)
8. Cui, H., et al.: Co-graph attention reasoning based imaging and clinical features integration for lymph node metastasis prediction. In: Proc. Int. Conf. MICCAI (pp. 657–666). Springer, Cham (2021)
9. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proc. ICML, pp. 8748–8763. PMLR (2021)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T. In Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, PMLR, pp. 4904–4916 (2021)
11. Ding, M., et al.: Cogview: mastering text-to-image generation via transformers. Adv. Neural Inf. Process. Syst. **34**, 19822–19835 (2021)

12. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F. In Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision, pp. 121–137. Springer, Berlin (2020)
13. Chen, Y.C., et al.: UNITER: UNiversal Image-TExt Representation Learning. In: Proc. ECCV, pp. 104–120. Springer, Cham (2020)
14. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: In Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2021)
15. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., Parikh, D. In Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
16. Li, J., et al.: Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. **34** (2021)
17. Kim, W., et al.: ViLT: vision-and-language Transformer without convolution or region supervision. In: ICML, pp. 5583–5594. PMLR (2021)
18. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)
19. El-Sappagh, S., et al.: Ultimodal multitasks deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing **412**, 197–215 (2020)
20. Holste, G., et al.: End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In: Proc. ICCV, pp. 3294–3303. IEEE (2021)
21. Yan, R., et al.: Richer fusion network for breast cancer classification based on Multi-modal data. BMC Med. Inform. Decis. Mak. **21**(1), 1–15 (2021)
22. Yao, J., et al.: Deep correlational learning for survival prediction from multi-modality data. In: Proc. MICCAI, pp. 406–414. Springer, Cham (2017)
23. Cheerla, A., et al.: Deep learning with multi-modal representation for pan-cancer prognosis prediction. Bioinform. **35**(14), 446–454 (2019)
24. Araújo, T., et al.: Classification of breast cancer histology images using convolutional neural networks. PLoS ONE **12**(6), e0177544 (2017)
25. Schulz, S., et al.: Multi-modal deep learning for prognosis prediction in renal cancer. Front. Oncol. **11** (2021)
26. Lu, M.Y., et al.: AI-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021)
27. Mobadersany, P., et al.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. USA **115**(13), E2970–E2979 (2018)
28. Yap, J., et al.: Multi-modal skin lesion classification using deep learning. Exp. Dermatol. **27**(11), 1261–1267 (2018)
29. Kawahara, J., et al.: Seven-point checklist and skin lesion classification using multitask multi-modal neural nets. IEEE J. Biomed. Health. Inform. **23**(2), 538–546 (2018)
30. Yoo, Y., et al.: Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from the clinically isolated syndrome. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **7**(3), 250–259 (2019)
31. Guan, Y., et al.: Predicting esophageal fistula risks using multi-modal self-attention network. In: Proc. Int. Conf. MICCAI, pp. 721–730. Springer, Cham (2021)
32. Silva, L., et al.: Pan-cancer prognosis prediction using multi-modal deep learning. In: Proc. ISBI, pp. 568–571. IEEE (2020)
33. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. **12**(6), 1247–1283 (2000)
34. Braman, N., et al.: Deep orthogonal fusion: Multi-modal prognostic biomarker discovery integrating radiology, pathology, genomics, and clinical data. In: Proc. MICCAI, pp. 667–677. Springer, Cham (2021)
35. Chen, R.J., et al.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans. Med. Imaging **41**(4), 757–770 (2022)
36. Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M.: Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci. Rep. **9**(1), 16884 (2019)
37. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)
38. Ramesh, A., et al.: Zero-shot text-to-image generation. In: Proc. ICML, pp. 8821–8831. PMLR (2021)
39. Li, J., et al.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086 (2022)
40. Huo, Y., et al.: WenLan: Bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561 (2021)
41. Hu, R., et al.: Unit: Multi-modal multitask learning with a unified Transformer. In: Proc. ICCV, pp. 1439–1449. IEEE (2021)
42. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: Proc. MICCAI, pp. 589–599. Springer, Cham (2021)
43. Li, S., et al.: A novel pathological images and genomic data fusion framework for breast cancer survival prediction. In: Proc. Int. Conf. EMBC, pp. 1384–1387. IEEE (2020)
44. Zhou, J., et al.: Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction. IEEE Trans. Circuits Syst. Video. Technol. (2021)
45. Li, H., et al.: Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In: Proc. MICCAI, pp. 529–539. Springer, Cham. (2021)
46. Zhou, T., et al.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum. Brain. Mapp. **40**(3), 1001–1016 (2019)
47. Li, X., et al.: Self-supervised feature learning via exploiting multimodal data for retinal disease diagnosis. IEEE Trans. Med. Imaging **39**(12), 4023–4033 (2020)
48. Floridi, L., Chiriatti, M.J.M.: Machines, GPT-3: its nature, scope. Lim. Conseq. **30**(4), 681–694 (2020)
49. Xu, Z., So, D., Dai, A.: MUFASA: Multi-modal fusion architecture search for electronic health records. Proc. AAAI Conf. Artif. Intell. **35**(12), 10532–10540 (2021)
50. Adams, J.: Information and misinformation in bibliometric time-trend analysis. J. Infometr. **12**(4), 1063–1071 (2018)
51. Garfield, E.: From the science of science to Scientometrics: visualizing the history of science with HistCite software. J. Informetr. **3**(3), 173–179 (2009)
52. Chen, C.: CityPlace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Assoc. Inf. Sci. Technol. **57**(3), 359–377 (2006)
53. Chen, C.: Searching for intellectual turning points: Progressive knowledge domain visualization. Proc. Natl. Acad. Sci. **101**(Suppl. 1), 5303–5310 (2004)
54. Van, E.N.J., et al.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010)
55. Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D.: Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage **145**, 137–165 (2017)
56. Shin, H.C., et al.: Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D

patient data. IEEE Trans. Pattern Anal. Mach. Intel. **35**(8), 1930–1943 (2012)

57. Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y.: A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med. Image Anal. **43**, 98–111 (2018)
58. Mahmud, M., Kaiser, M.S., Hussain, A., Vassanelli, S.: Applications of deep learning and reinforcement learning to biological data. IEEE Trans. Neural. Netw. Learn. Syst. **29**(6), 2063–2079 (2018)
59. Estai, M., Bunt, S.: Best teaching practices in anatomy education: A critical review. Ann. Anat. **208**, 151–157 (2016)
60. Liu, S., et al.: Multi-modal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. **62**(4), 1132–1140 (2014)
61. Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G.: Representation learning for mammography mass lesion classification with convolutional neural networks. Comput. Methods. Programs Biomed. **127**, 248–257 (2016)
62. Liu, Z., et al.: The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges. Theranostics **9**(5), 1303 (2019)
63. Bychkov, D., et al.: Deep learning-based tissue analysis predicts outcome in colorectal cancer. Sci. Rep. **8**(1), 1–11 (2018)
64. Zeng, X., et al.: deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinform. **35**(24), 5191–5198 (2019)
65. Fetit, A.E, et al. A multimodal approach to cardiovascular risk stratification in patients with type 2 diabetes incorporating retinal, genomic and clinical features. Sci. Rep. **9**(1), 3591 (2019)