Journal of Advances in Science and Technology
Vol. 21, Issue No. 1, March-2024, ISSN 2230-9659

DOI: https://doi.org/10.29070/24mw3w90

Check for updates

Multimodal Disease Classification and Severity analysis Approaches using Machine and Deep Learning

Manisha P. Patil [1] * , Dr. Uruj Jaleel [2]

1. PhD Scholar, Kalinga University, Raipur, C.G. , India
bhagat.manisha0412@gmail.com ,
2. Associate Professor, Ph.D.Guide, Kalinga University, Raipur, C.G. , India

Abstract: Chest X-rays are a common diagnostic tool for pulmonary and cardiac conditions in hospitals because they provide a clear picture of the patient's thorax. With the use of image-to-text radiology report production, medical imaging results may be automatically described in radiology reports. There are a lot of different pieces of patient data that radiologists may access, but most current systems only use the picture data. the objective of developing AI systems with a focus on humans, with the ability to learn radiologists' search habits via their eye movements, with the hope of enhancing DL system categorisation. The goal of this research is to evaluate several multimodal DL architectures in collaboration with trained radiologists to see which ones work best. In particular, this study aims to build strong DL models for medical picture analysis by investigating the integration of several data modalities, such as eye tracking data and patients' clinical data. A multimodal DL model integrating clinical data and chest X-rays (CXRs) was suggested by us. Findings demonstrated that baseline performance was unaffected by directly supplying fixation masks of radiologists' gaze patterns as input. Confine Areas Using R-CNN (Recurrent Neural Networks).

Keywords: Multi-modal, Deep Learning, Chest, Radiology, X-ray

X

INTRODUCTION

Medical imaging is extensively used in many areas of the health sciences for the purpose of illness diagnosis, treatment plan development, patient care, and outcome prediction. Based on their findings and other pertinent clinical data and information, including patient demographics, symptoms, and pre-existing/existing medical problems, radiologists are tasked with interpreting medical pictures and generating a full-text radiology report. In addition to being thorough and precise, these reports also need to be prepared quickly in a certain manner. When diagnosing lung disorders in clinical settings, the first step in the evaluation process is often a chest x-ray (CXR), the most used medical imaging tool. Typical CXR reports contain "findings" and "impressions" sections where radiologists note what they think are normal and aberrant aspects of the pictures. Writing these in-depth reports is not only a challenging and error-prone task, but also demands a great deal of expertise and experience. Robots may cut down on mundane tasks by giving radiologists a baseline to check and adjust as necessary. As a result, radiologists would have more time for advanced clinical reasoning and ensuring patient safety.

Automated production of radiology reports describing medical image results is the goal of image-to-text radiology report creation. The majority of current approaches ignore all other patient data available to radiologists in favour of analysing images alone. Our innovative multi-modal deep neural network

architecture incorporates both organised patient data, such vital signs and symptoms, and unstructured clinical notes to produce chest x-ray results. There have been efforts to automate the production of radiology reports in the area of medical imaging informatics [1, 2]. Figure 1 shows that most existing deep learning methods rely on networks initially developed for picture captioning, which include a convolutional encoder and a recurrent or transformer decoder.

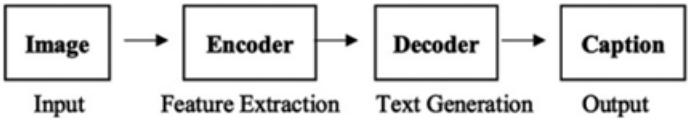

Figure 1: A framework for generalising image-to-text

Despite the fact that the input and output modalities are identical across the two tasks, there are significant distinctions. Comprehensive and including particular medical facts, radiology reports are written in the form of full paragraphs instead of simple captions. In addition, there are frequently little differences between the picture and the report, which makes it difficult to understand medical images. Moreover, in order to provide a description of a medical image, additional information beyond what is seen in the image is typically required. For example, in certain instances, when comparing medical imaging between males and females, the visual patterns are almost same. However, when it comes to patient demographics, there are notable variances that might affect the evaluation and diagnosis. Present CXR report generating approaches, however, ignore the non-imaging information available to radiologists during image interpretation in favour of only using the radiological picture as input.

LITERATURE REVIEW

Altwijri, et al (2023) [3] This study's overarching goal is to construct a deep-learning strategy for AD severity level detection that makes use of pre-trained convolutional neural networks (CNNs), especially in cases where both the amount and quality of accessible datasets are constrained. Here, before training begins, the AD dataset is refined using an image processing module. Utilising four Kaggle AD datasets—one for the normal stage of the disease and three for the mild, very mild, and moderate stages, respectively—the suggested method was contrasted with two famous deep-learning algorithms (VGG16 and ResNet50). Because of this, we were able to assess how well the classification results worked. We compared the three models using six different performance metrics. Our method outperforms the competition with a detection accuracy of 99.3 percent, according to the results.

ZainEldin, et al (2022) [4] A convolutional neural network (CNN) hyperparameters optimisation approach called adaptive dynamic sine-cosine fitness grey wolf optimiser (ADSCFGWO) is used by the suggested Brain Tumour Classification Model based on CNN (BCM-CNN). After hyperparameter optimisation, an Inception-ResnetV2 training model is constructed. The model uses Inception-ResnetV2, a popular pre-trained model, to enhance brain tumour diagnosis. Its output is a binary number between 0 and 1, with 0 representing normal and 1 representing tumour. The ADSCFGWO algorithm is a flexible framework that takes use of the best features of both the sine cosine and grey wolf algorithms. Because the hyperparameters used for CNN optimisation improved the CNN's performance, the experimental results

demonstrate that the BCM-CNN classifier performed the best. Achieving an accuracy of 99.98% with the BRaTS 2021 Task 1 dataset was accomplished by the BCM-CNN.

Ieracitano, et al (2019) [5] This research presents a new method for automatically classifying brain states using EEG designed characteristics and multi-modal Machine Learning (ML). In order to distinguish between patients and Healthy Control (HC) individuals, electroencephalograms (EEGs) are recorded from neurological patients suffering from Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD). Extraction of higher-order statistics (HOS) from the bispectrum (BiS) representation is also done in order to take use of the nonlinear phase-coupling information found in EEG data. In addition to the five EEG sub-bands, BiS also produces a second set of characteristics called BiS features. Multiple machine learning classifiers use the CWT and BiS features to conduct 2-way (AD vs. HC, AD vs. MCI, MCI vs. HC) and 3-way (AD vs. MCI vs. HC) classifications. A balanced EEG dataset consisting of 63 AD, 63 MCI, and 63 HC is examined as an experimental benchmark. Based on the comparison results, the Multi-Layer Perceptron (MLP) classifier is the best option when using a combination of CWT and BiS features as input. The other models that performed better were Autoencoder (AE), Logistic Regression (LR), and Support Vector Machine (SVM). As a result, cutting-edge deep learning methods are computationally demanding, while our suggested multi-modal ML technique is computationally more efficient.

Venugopalan, et al (2021) [6] We integrate imaging (magnetic resonance imaging, or MRI), genetic (single nucleotide polymorphisms, or SNPs), and clinical test data using deep learning (DL) to categorise individuals into Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (CN). For clinical and genetic data, we use stacked denoising auto-encoders, and for imaging data, we employ 3D-convolutional neural networks (CNNs). With the use of the ADNI dataset, we show that deep models perform better than shallow ones, such as k-nearest neighbours, decision trees, random forests, and support vector machines. Better accuracy, precision, recall, and meanF1 scores are achieved by incorporating multi-modality data rather than single-modality models. The top distinguishing traits found by our models include the hippocampus and amygdala brain regions, as well as the Rey Auditory Verbal Learning Test (RAVLT), which aligns with the known AD literature.

Aparna, Mudiyala & Rao, Battula. (2023) [7] To address this, we use deep features retrieved from two deep learning models that have already been trained in this study. For the purpose of multi-class classification in Alzheimer's disease, the suggested models DenseNet121 and MobileNetV2 are used. At the beginning of this process, we used CycleGAN (generative adversarial networks) to produce pictures and expand the dataset by 70%. The suggested models were 98.82% accurate. The outcomes it produces are superior than those of competing models.

RESEARCH METHODOLOGY

A Deep Neural Architecture with Multiple Modalities

Figure 2 shows the proposed design of the neural network with two independent branches that take X-ray pictures and the related radiologist reports as inputs and produce meaningful representations. There are three primary parts to the text classification branch: a multi-head neural attention [8] mechanism, Bidirectional Multiplicative Long Short-Term Memory (bi-mLSTM) units, and pretrained biological word

embeddings. Next, state-of-the-art convolutional neural network (CNN) models like ResNet, DenseNet, DPN, or EfficientNet are used by the image classification subfield. A 14-node prediction layer is informed by the characteristics that are obtained when the representations from both branches are combined in the entire network. The sigmoid activation function was used to transform each node's score between 0 and 1, irrespective of the other nodes' values, since the process of categorising radiological tests involves several labels.

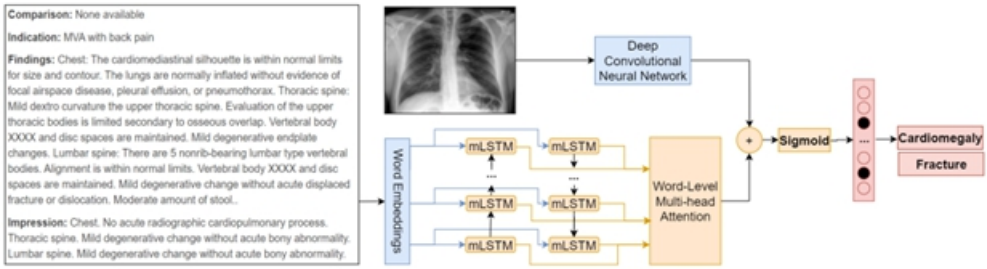

Figure 2: An outline of the neural network design that has been suggested. The blue boxes show the parts of our overall design that we put various pre-trained methods to the test

Textual Data Sources

We investigate BioWordVec, BioELMo [9], and BioBERT, three pre-trained word embeddings, for text representation. Pre-trained across massive biomedical datasets, including MIMIC-III clinical notes and PubMed abstracts, are BioWordVec FastText word embeddings. Predicting neighbouring words from a centre word is the goal of the skipgram model. The approach aims to maximise the following average log probability for a series of words w1, w2,..., wN, representing a textual corpus N words, and a c context windows size:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{-c\leq i\leq c, i\neq 0}\log\left(\mathrm{p}\left(w_{n+i}|w_n\right)\right)\right)$$

.....(1)

Where the word at position n is represented by wn. The following equation technically describes the usual definition of p (wn+i |wn) as it applies a softmax function:

$$\mathrm{p}\left(w_I|w_O\right)=\frac{\exp\left(\mathrm{V}\left(w_I\right)^{\top}\mathrm{V}\left(w_O\right)\right)}{\sum_{w=1}^{W}\exp\left(\mathrm{V}(w)^{\top}\mathrm{V}\left(w_O\right)\right)}$$

.....(2)

This is where W stands for the vocabulary size, V (wI) for an input centre word and V (wO) for an output context word are vector representations, respectively. An n-gram of characters represents each word in the FastText word embedding approach, which is an extension of the Skip-gram model.

In the pre-training corpus, the 1 million most common tokens make up BioELMo's lexicon. Specifically, the RNN architecture makes use of multiply LSTMS (mLSTMs) and stacked layers of bidirectional RNNs

that use long short-term memory (LSTM) cells—a specific sort of RNN that will be described later on. The 2048 channel char-ngram CNN, as shown in Figure 3, generates the context-independent token representation. Then, two highway layers—an expansion of the residual connections concept—modulate the amount of input signal to be added to the output. The input is passed through fully connected layers with sigmoid and ReLU as non-linearity activation functions, and finally, a linear projection is applied down to 512 dimensions. BioBERT makes use of a model that was previously trained on generic texts found in PubMed abstracts and PubMed central full-text articles [10], specifically the Bidirectional Encoder Representations from Transformers (BERT) model, which was developed by Devlin et al. (2018). An alternative neural design called the Transformer [8] is used by BERT to represent word sequences instead of RNNs. This architecture consists of a stack of encoder and decoder blocks.

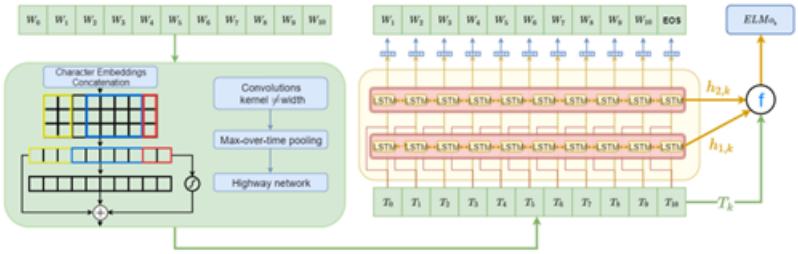

Figure 3: ELMo's design showcases how contextual embeddings may be created with BiLM

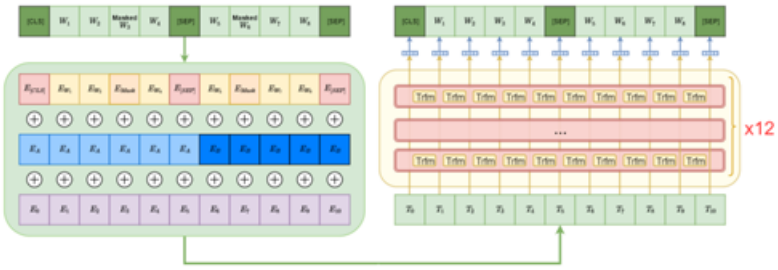

Figure 4: Designing contextual embeddings using the MLM task is shown via BERT's architecture

Figure 4 shows the segmented sequence in action, beginning with the [CLS] token and continuing with the WordPiece tokens divided into two sentences each by two [SEP] tokens. The input embeddings in a Transformer architecture are the total of the embeddings for tokens, segmentation, and positions.

Pre-Training for the Per-Modality Model

The suggested multi-modal architecture did not start with arbitrarily initialised parameters; instead, the text and image processing routes were pre-trained on huge datasets from a single modality. More specifically, each of the CNN architectures mentioned was pre-trained and evaluated using a randomly shuffled dataset from MIMIC-CXR and CheXpert.

The Multimodal Eye Imaging (MIMIC-EYE) Database

Medical pictures, reports, clinical data, eye tracking data, gaze, and pupil dilation information are all part

of MIMIC-EYE [11], a complete dataset that we developed for this article. The visual search habits of radiologists and the construction of DL models might be better understood with the combination of eye tracking data with MIMIC modalities. There are 3,689 tuples in MIMIC-Eye that comprise chest X-ray pictures, eye-gaze data, and voice transcripts from radiologists. Among them, 1,683 tuples include clinical data.

DATA ANALYSIS

For our experiments, we used four datasets: two sets of frontal chest X-ray images annotated with 14 observation classes (CheXpert and the original MIMIC-CXR dataset), one set of full-text radiology reports annotated with the same 14 observation classes (a subset of MIMIC-III data with ICD labels converted to these same classes), and lastly, the Open-i multimodal radiography dataset annotated with the same 14 observation classes annotated with frontal chest X-ray images and full-text reports from multiple locations collected by Indiana University, along with the same fourteen observation classes. General statistics for data characterisation are shown in Table 1.

Table 1: An analysis of the experimental datasets from a statistical perspective

Label	Images		Text	Multi-modal
	MIMIC-CXR	CheXpert	MIMIC-III	Open-i
No Finding	83,336	19,765	133,563	2,062
Cardiomediastinum	18,240	19,578	608	0
Cardiomegaly	56,012	30,158	351	323
Lung Opacity	60,196	98,759	62	417
Lung Lesion	8,315	8,149	9,741	1,291
Edema	43,812	61,535	4,635	67
Consolidation	16,614	37,396	43,006	28
Pneumonia	38,262	20,664	82,526	78
Atelectasis	61,108	59,658	287	349
Pneumothorax	12,953	20,408	16,241	25
Pleural Effusion	65,449	86,541	29,978	160
Pleural Other	3,009	4,318	896	37
Fracture	5,675	7,935	16,862	97
Support Devices	74,970	108,184	10,504	48
Total Instances	250,044	191,229	261,091	3,689

Specifically, between 2011 and 2016, the MIMIC-CXR dataset has 371,920 chest X-rays linked to 227,943 investigations involving patients hospitalised to the Beth Israel Deaconess Medical Centre. Conversely, CheXpert is comprised of 65,240 patients' chest X-rays totalling 224,316, acquired from Stanford Hospital from October 2002 to July 2017. When conducting experiments using image data, such as when comparing different CNN architectures or when pre-training the entire model for multi-modal tests, we exclusively used frontal view X-ray images (i.e., 250,044 instances from MIMIC-CXR and 191,229 instances from CheXpert). From patient discharge notes, MIMIC-III compiles radiological reports linked to ICD diagnosis codes; the database is available to the public and is used for critical care purposes.

After matching the 14 labels in the MIMIC-CXR dataset with the corresponding set of ICD codes, we used these to filter the radiology reports based on the frequency of key-phrases such "chest," "lungs," and "thorax." Also, we used these for our analysis. The end product was a collection of 261,091 text documents.

Radiography Data Sorting Assignment

The tests took into account two distinct configurations for the radiography chest X-ray picture categorisation task: (i) classify X-ray greyscale images trained with the combined MIMIC-CXR/CheXpert data using the DPN-92, ResNet-50, DenseNet-101, and EfficientNet-B5 architectures; and (ii) classify X-ray RGB images trained with the combined MIMIC-CXR/CheXpert data using the ResNet-50, DenseNet-101, and EfficientNet-B5 models pre-trained with ImageNet.

Table 2: Findings from experiments conducted using the MIMIC-CXR/CheXpert dataset

Model	Accuracy	LRAP	CE	Precision		Recall		F1-score	
				Micro	Macro	Micro	Macro	Micro	Macro
DPN-92	0.8173	0.6032	6.2742	0.1105	0.0727	0.5157	0.1701	0.1821	0.0693
DenseNet-201	0.8285	0.5995	6.3483	0.2363	0.1300	0.5837	0.3274	0.3364	0.1306
ResNet-50	0.8332	0.6115	6.3311	0.3972	0.2070	0.5664	0.2890	0.4671	0.1921
EfficientNet-95	0.8466	0.6873	5.5841	0.3061	0.1888	0.6861	0.3695	0.4233	0.2266
ResNet-50 w\ ImageNet pre-train	0.8434	0.6441	5.9881	0.3894	0.2071	0.6181	0.3251	0.4777	0.2093
DenseNet-201 w\ ImageNet pre-train	0.8465	0.6866	5.6044	0.2914	0.1655	0.6985	0.3686	0.4113	0.1863
EfficientNet-B5 w\ ImageNet pre-train	0.8636	0.7247	5.2661	0.4528	0.2733	0.6997	0.4868	0.5498	0.3074

You can see the outcomes of the testing data divides in Table 2. Setting (i) is represented by the first four rows, and Setting (ii) by the next three rows. Results from the first experimental setting showed that EfficientNet-B5 performed better across the board for almost all measures, while DPN-92 performed worse across the board for nearly all metrics. In the second experimental setting, we found that using publicly available ImageNet pre-trained weights instead of randomly initialised ones and using RGB data instead of greyscale or other 1-channel image data improved the models' overall performance. For example, when comparing results with settings (ii) and (i), EfficientNet-B5 consistently outperformed with settings (ii). Figure 5 shows the AUROC values for each class for each model, which gives more information. With micro and macro average AUROC scores of 0.869 and 0.792, respectively, EfficientNet-b5 surpassed all other models.

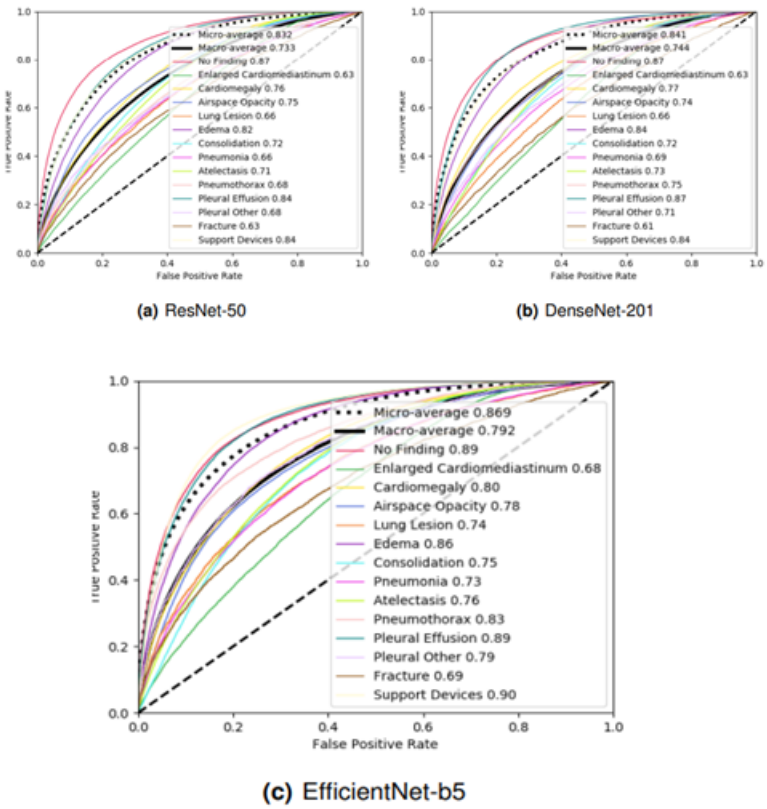

Figure 5: Visual representations of data showing how well the experimental environment classified

MDF-Net: A Network for Multimodal Dual-Fusion

Following the creation of the MIMIC-Eye dataset, a novel architecture known as multimodal Dual-Fusion Network (MDFNet) was designed. This network combines clinical data with CXR pictures to identify lesions. The disparity in dimensions between the two modalities is a major obstacle to their effective use; for example, clinical data is often represented as a one-dimensional tensor, whereas photographs are usually three-dimensional. We solved this problem by introducing spatialization, a method that transforms two-dimensional clinical data into three-dimensional space. To enable the fusing of pictures and clinical data in three-dimensional space, our spatialization module has several deconvolutional layers that increase the pixel size to the required level for fusion. Lesion detection performance and model generalisation were both enhanced by 12% on Average Precision (AP) when clinical data was included, as shown using MDF-Net. Table 3 shows that our innovative architecture and spatialization technique provide a potential way to use deep learning with multimodal data for medical picture analysis.

Table 3: Findings from the planned MDF-Net evaluation

Lesion	Mask R-CNN (Baseline)		MDF-Net (3D Fusion only)		MDF-Net	
	AP	AR	AP	AR	AP	AR
Enlarged Cardiac Silhouette (Enl. Card. Sil.)	54.82	100.01	69.11	94.43	70.35	101.00
Atalectasis	14.57	41.01	16.32	42.85	24.42	48.56
Pleural Abnormality (Pleural Abn.)	16.28	42.85	13.21	52.37	16.08	28.56
Consolidation	3.14	20.04	16.92	40.04	14.28	30.04
Pulmonary Edema (Pulm. Edema)	9.21	66.66	22.25	72.21	33.24	66.66
Overall	19.60	53.91	27.56	60.37	31.68	54.75

The dual fusion method achieves better performance than the baseline MaskRCNN by around 12% AP and the MDF-Net(3D) by 4.12% when using a score threshold of 0.05 and an IoBB threshold of 0.5, according to the results. Table entries for enlarged cardiac silhouette, abnormalities of the pleura, and pulmonary oedema are as follows: Enl. Card. Sil., Pleural Abn., and Pulm. Oedema.

CONCLUSION

In conclusion, radiologists may benefit from deep learning when it comes to CXR interpretation. In order to identify lesions in CXRs, this study looks at how to incorporate eye tracking data into DL structures. The first findings demonstrated that the baseline Mask RCNN was not improved by explicitly giving fixation masks of radiologists' gaze patterns as input. The great performance of the multi-modal end-to-end architecture's constituent components was verified. Particularly important to the overall model performance was pre-training. The findings show that how the models are trained beforehand really affects how well they operate. The suggested method made use of pre-trained weights from the publicly available dataset ImageNet, which does not include any medical imaging data. The results on the challenge may suggest that the learnt characteristics for X-ray image classification are comparable to those for everyday picture classification. Investigating MRI image reconstruction also made use of the same dataset. In the future, researchers will try to solve the problem of DL methods that generate noisy data from raw eye gaze data without taking the complexities of human creation into account, which makes the data unusable for supervised learning. Train on the downstream job of categorising chest radiology data to fine-tune both BioBERT and BioELMo, instead of utilising them as fixed feature extractors. With a fraction of the dimensions of state-of-the-art Transformer models, this method has shown promise in language modelling problems.

References

1. Yang S, Niu J, Wu J, Liu X. (2020). Automatic medical image report generation with multi-view, multi-modal attention mechanism. In: International Conference on Algorithms, Architectures for Parallel Processing. Springer; p. 687–99.

2. Singh S, Karimi S, Ho-Shon K, Hamey L. (2021). Show, tell and summarise: learning to generate and

summarise radiology findings from medical images. Neural Comput Appl. 33:7441–65. 10.1007/s00521-021-05943-6

3. Altwijri, O.; Alanazi, R.; Aleid, A.; Alhussaini, K.; Aloqalaa, Z.; Almijalli, M.; Saad, A. (2023) Novel Deep-Learning Approach for Automatic Diagnosis of Alzheimer's Disease from MRI. Appl. Sci., 13, 13051. https://doi.org/10.3390/app132413051

4. ZainEldin, H., Gamel, S. A., El-Kenawy, E. M., Alharbi, A. H., Khafaga, D. S., Ibrahim, A., & Talaat, F. M. (2022). Brain Tumor Detection and Classification Using Deep Learning and Sine-Cosine Fitness Grey Wolf Optimization. Bioengineering (Basel, Switzerland), 10(1), 18. https://doi.org/10.3390/bioengineering10010018

5. Ieracitano, Cosimo & Mammone, Nadia & Hussain, Amir & Morabito, Francesco. (2019). A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. Neural networks : the official journal of the International Neural Network Society. 123. 176-190. 10.1016/j.neunet.2019.12.006.

6. Venugopalan, Janani & Tong, Li & Hassanzadeh, Hamid Reza & Wang, May. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific Reports. 11. 3254. 10.1038/s41598-020-74399-w.

7. Aparna, Mudiyala & Rao, Battula. (2023). A novel automated deep learning approach for Alzheimer's disease classification. IAES International Journal of Artificial Intelligence (IJ-AI). 12. 451. 10.11591/ijai.v12.i1.pp451-458.

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Proceedings of the Neural Information Processing Systems.

9. Jin, Q., Dhingra, B., Cohen, W., and Lu, X. (2019). Probing biomedical embeddings from language models. In Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP.

10. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

11. C. Hsieh et al. (2023) MIMIC-Eye: Integrating MIMIC Datasets with REFLACX and Eye Gaze for Multimodal Deep Learning Applications. PhysioNet.