SGMoE: Semantics-Guided Mixture of Experts for Multi-class Disease Classification Based on Radiology Reports

Anonymous ACL submission

Abstract

Radiology reports contain detailed image description that is crucial for clinicians in decision-making, and automated disease classification based on radiology reports can be more effective than image-based classification. Although deep learning achieves promising performance on this task, existing approaches struggle with increasing class complexity when multi-class disease classification is considered, where multiple distinct modes can coexist within the feature distribution of a single class, leading to highly complicated decision hypersurfaces. To address this challenge, we propose Semantics-Guided Mixture of Experts (SGMoE) for report-based multi-class disease classification. SGMoE specializes multiple classification experts in handling different disease modes, each learning class boundaries within a specific subspace of the feature distribution. To guide the subspace allocation for each expert, SGMoE uses the report semantics to determine the expert assignment. This is achieved by clustering the report semantic embedding, and then an expert is assigned to determine specific classes in a certain cluster or clusters. Moreover, a gating network is designed to adaptively select appropriate experts for final classification, with a gating loss penalizing gating that contradicts with the expert assignment for model training. Experiments on an in-house dataset of 11,864 reports and the public CT-RATE dataset show that SGMoE achieves more accurate multi-class disease classification than existing text classification approaches.

1 Introduction

Automated disease classification based on radiology images, e.g., magnetic resonance imaging (MRI), computed tomography (CT), and X-ray imaging, has drawn considerable attention for its potential to aid clinicians in making timely and precise diagnoses, ultimately improving patient outcomes (Wolbarst and Hendee, 2006). In particular, deep learning (DL) methods have led to promising progress in the classification task. For example, convolutional neural network (CNN) models have been applied to brain tumor classification, cardiac disease detection, and chest radiograph diagnosis (Akkus et al., 2016; Choi et al., 2021; Korfiatis et al., 2017; Xu et al., 2023; Rajpurkar et al., 2018). Similarly, Transformer-based models have also demonstrated superior performance in tumor molecular status prediction, breast cancer classification, and chest X-ray diagnosis (Chen et al., 2024; Wang et al., 2022; Park et al., 2022). Despite these advancements, it can still be challenging for these image-based methods to achieve high classification accuracy, as the image dimension is usually high and it is difficult to learn an optimal association between disease types and image appearances (Litjens et al., 2017; Zhang et al., 2019).

Recent studies have explored disease classification based on radiology reports (Chen et al., 2018; Kim et al., 2019; Tang et al., 2020), and it is shown in (Gao et al., 2024) that report-based disease classification can be more accurate than image-based classification. In clinical practice, radiology reports written by radiologists accompany radiology images, and they provide detailed descriptions about the important observations, e.g., anomalies, in the image. As report drafting is required by clinical routine, report-based classification does not introduce additional human workload. In the report-based disease classification, the findings part that describe the image appearance is used without the impressions part that give potential diagnostic suggestions, as it is easier for radiologists to describe the image than determine the disease type. Even less experienced radiologists can provide high-quality image description, but they are often unable to diagnose complicated diseases. The report information has a much lower dimension, which makes it easier to learn the association between image appearances and disease

types. The report-based classification in (Gao et al., 2024) encodes the text information with Bidirectional Encoder Representations from Transformers (BERT) model variants (Devlin, 2018), and the classification accuracy is shown to be clearly better than image-based models, including ResNet (He et al., 2016), ViT (Dosovitskiy, 2020), Swin Transformer (Liu et al., 2021), etc.

The report-based disease classification may still struggle to address multi-class scenarios when the distribution of certain disease types can have multiple distinct modes in the feature space. For example, different types of brain tumors may share overlapping semantic or imaging characteristics, while simultaneously displaying significant variations within the same type (Mirbabaie et al., 2021). The multimodal distribution increases the difficulty of finding a single classification boundary to discriminate the disease (Ruff et al., 2021), and it becomes even more challenging when there are a considerable number of different disease types to classify (Bilal et al., 2017).

To address the challenge of multimodal distribution in multi-class disease classification, we seek to decompose the complex classification task into smaller and more manageable subproblems. Accordingly, in this work we propose the Semantics-Guided Mixture of Experts (SGMoE) framework for multi-class classification based on radiology reports.

In SGMoE multiple experts are used, each specializing in a specific subset of the problem space. In this way, instead of finding a single set of complex classification hypersurfaces, SGMoE determines multiple simple decision boundaries, greatly simplifying the classification task. To determine the assignment of the subspace to each expert, SGMoE develops a semantics-guided clustering strategy, which explores the inherent structure of the report data. In particular, the semantic embeddings of the reports are clustered to partition the feature space into different subspace. After denoising the potential data noise, each cluster only comprises a subset of all disease types, and each expert is dedicated to classifying the diseases within one or a few clusters. Note that when a disease type has multiple modes, different modes can be separated in different clusters and handled by different experts, and the expert is no longer faced with the challenging issue of multimodal feature distribution. Given the class assignment of each expert, to determine how experts are activated for an input sample, SGMoE employs a gating network to dynamically feed the input to appropriate experts. The gating network considers both the input information and prior cluster information.

To jointly train the experts and gating network, SGMoE further designs a gating loss that penalizes expert activations that contradict with the class assignment, and it is used together with a standard classification loss for model training. To validate SGMoE, experiments were performed on a private dataset and the public CT-RATE dataset (Hamamci et al., 2024a,b,c), where SGMoE outperforms competing text-based classification models. The SGMoE code will be publicly available.

2 Related Work

2.1 DL Disease Classification Based on Radiology Data

DL models have been widely applied to the diagnosis of various diseases, including neurological diseases, chest diseases, heart diseases, etc. Conventionally, radiology images such as MRI, CT, and X-ray scans are used for disease classification. For example, CNN-based models have been developed to classify gliomas based on MRI by predicting the isocitrate dehydrogenase mutation status, which plays a crucial role in treatment planning (Choi et al., 2021). Additionally, CNNs have been employed for the early diagnosis and classification of Alzheimer's disease using MRI (Salehi et al., 2020). In the context of chest diseases, CNN-based models have shown high performance in identifying pneumonia and diagnosing COVID-19 infections from chest X-ray (Islam et al., 2020). Similarly, in cardiovascular diseases, CNN-based models have been utilized to detect coronary artery disease and other heart abnormalities on CT scans (Xu et al., 2023). Beyond CNN-base models, Transformer-based models have emerged as a promising alternative, demonstrating superior performance in tasks such as tumor molecular status prediction on MRI (Chen et al., 2024), breast cancer classification on mammography images (Abimouloud et al., 2024), chest diseases diagnosis on X-ray scans (Park et al., 2022), etc. However, radiology images usually have a high dimension, and the learning of effective feature extraction from the high-dimensional data for disease classification can be challenging, which limits the classification performance for more difficult tasks.

In addition to radiology images, several existing

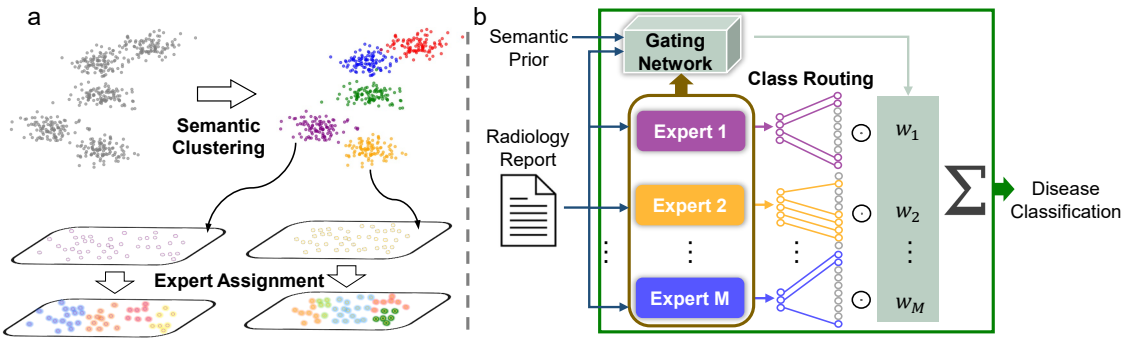

Figure 1: An overview of SGMoE: (a) semantic clustering of training data for expert assignment; (b) the overall architecture of SGMoE given the expert assignment.

works have explored the use of the findings part of radiology reports for disease classification, with the expectation that the lower dimension of the report information can reduce the difficulty in learning the association between image appearances and disease types. For example, natural language processing (NLP) algorithms powered by machine learning have demonstrated success in identifying acute ischemic stroke phenotypes from brain MRI reports (Kim et al., 2019). CNN-based models have been applied to classify pulmonary embolism findings from thoracic CT reports, achieving high accuracy and outperforming traditional NLP methods (Chen et al., 2018); they have also been used to classify normal and abnormal chest radiographs based on the findings part of the reports, with remarkable diagnostic accuracy and high generalizability (Tang et al., 2020). Furthermore, in (Gao et al., 2024), BERT-based language models are fine-tuned for disease classification, and it is observed that report-based classification is on average about 10% more accurate than image-based models.

These works show the promising potential of using radiology reports for disease classification and underscore the importance of advancing text-based models in medical diagnosis.

2.2 Text Classification

Report-based disease diagnosis is closely related to text classification, which is a foundational task in NLP. Its applications include sentiment analysis, topic modeling, spam detection, etc. Traditional approaches to text classification typically rely on hand-crafted features and shallow machine learning algorithms, for example, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017).

With the development of DL techniques, text classification performance has been drastically improved.

Early advancements in this area are driven by recurrent neural networks (RNNs), particularly the long short-term memory (LSTM) network, which effectively captures sequential dependencies in text (Yu et al., 2019). The LSTM network is useful for processing medical reports with detailed contextual information (Liu et al., 2022).

More advanced Transformer architectures are also used for text classification, and the BERT model (Devlin, 2018) is a notable example. BioBERT (Lee et al., 2020), RadBERT (Yan et al., 2022), and ClinicalBERT (Huang et al., 2019) extend BERT to biomedical and clinical domains based on domain-specific corpora such as PubMed abstracts and electronic health records, and they achieve superior performance in medical text analysis. RoBERTa (Liu, 2019) improves upon BERT with dynamic masking during pretraining, leading to enhanced contextual understanding, and it may be extended to the medical domain as well.

3 Method

3.1 Method Overview

An overview of the SGMoE framework that classifies multiple disease types based on radiology reports is shown in Fig. 1. Instead of addressing the multi-class problem with one single classifier, SGMoE uses different classification experts that are responsible for different subsets of diseases, which decomposes the challenging multi-class classification task into manageable subproblems. SGMoE first uses semantic clustering to determine the expert assignment of disease classes as illustrated in Fig. 1a.

Then, the overall architecture of SGMoE given the expert assignment is shown in Fig. 1b. SGMoE

comprises multiple experts determined by the semantic clustering, and they are followed by class routing to merge the classification results for different disease types. A gating network is used to allocate the input sample to appropriate experts, where the gating is decided by both prior knowledge from semantic clustering and input-specific features. The detailed design of SGMoE is presented below.

3.2 Expert Assignment Guided by Semantic Clustering

Before constructing the network of SGMoE, the specific roles of the experts should be determined. SGMoE seeks to partition the feature space so that each expert excels in the classification of a subset of disease types within a subspace, where multimodal disease distributions can be decomposed into multiple simple distributions handled by different experts. To this end, a pretrained RoBERTa model (Cui et al., 2020, 2019) is used to obtain the semantic embeddings of radiology reports, and semantic clustering is performed based on the embeddings for expert assignment. Specifically, the simple k-means clustering algorithm (Jain and Dubes, 1988) is applied to the training reports, for which the disease type is known, and the optimal number N of clusters is determined based on the silhouette coefficient (Rousseeuw, 1987).

As the semantic embeddings may contain noise and/or inaccuracy, each cluster can comprise disease types with a very small number of samples, and directly assigning all classes in the cluster to the same classification expert may lead to high class imbalance. Therefore, expert assignment is performed after class denoising, where for the j-th cluster ($j \in \{1, \ldots, N\}$), the corresponding set $\mathcal{S}_j$ of assigned classes only include those with a sufficient number of samples:

$$\mathcal{S}_j = \left\{ a_i \mid \frac{n_{i,j}}{n_i} \geq \frac{1}{N} \right\}. \tag{1}$$

Here, a_i denotes the i-th disease type, $n_{i,j}$ is the total number of training samples of a_i, and $n_{i,j}$ is the number of training samples of a_i in the j-th cluster. Eq. (1) removes samples that are clustered likely due to noise and also ensures that each class appears at least in the assignment of one cluster.

Note that some clusters can only contain one single disease type after the denoising, and the classification becomes trivial for the expert. To address the problem, the classes in those clusters with a single class are grouped and assigned to a single expert, as the samples in these clusters are compact in the feature space and easy to separate.

3.3 SGMoE Architecture

The expert assignment guided by semantic clustering leads to M ($M \leq N$) experts, where the m-th expert E_m handles the class set $\mathcal{C}_m$.

For an input sample x, E_m outputs the classification probability vector $\boldsymbol{p}_m = [p_{m,1}, p_{m,2}, \ldots, p_{m,|\mathcal{C}_m|}]$ for its assigned classes, where $|\mathcal{C}_m|$ is the cardinality of $\mathcal{C}_m$. As it is possible that multiple experts may be activated for a sample, the results of different experts need to be fused. However, different $\boldsymbol{p}_m$'s are associated with different combinations of classes, and they cannot be directly fused.

To address the problem, we first develop a class routing mechanism as follows for expert fusion. Specifically, $\boldsymbol{p}_m$ is mapped back to a vector of length D, where D is the number of disease types of interest, with a permutation matrix $\mathbf{P}_m \in \mathbb{R}^{|\mathcal{C}_m| \times D}$ as

$$\boldsymbol{p}'_m = \boldsymbol{p}_m \mathbf{P}_m. \tag{2}$$

The entries $P_m^{k,i}$ in $\mathbf{P}_m$ are determined as

$$P_m^{k,i} = \begin{cases} 1, & \text{if } a_i \in \mathcal{C}_m \text{ and } \mathrm{index}(a_i, \mathcal{C}_m) = k \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathrm{index}(a_i, \mathcal{C}_m)$ represents the index of a_i in $\mathcal{C}_m$. In this way, all expert outputs are reshaped as D-length vectors $\boldsymbol{p}'_m$, and their entries are now aligned, with non-assigned classes having zero values.

To fuse the expert outputs, we develop a gating network. It gives a weight vector $\boldsymbol{w} = (w_1, \ldots, w_M)$ and aggregates $\boldsymbol{p}'_m$'s as

$$\boldsymbol{p}_{\text{final}} = \sum_{m=1}^{M} w_m \boldsymbol{p}'_m, \tag{4}$$

where $\boldsymbol{p}_{\text{final}}$ is the final classification result. This aggregation ensures that the final classification decision integrates the specialized knowledge of all experts while prioritizing the contributions of the most relevant experts based on the gating output.

Unlike standard mixture of experts, the gating here considers prior knowledge from semantic clustering in addition to input data, so that expert activation both conforms to the semantic priors and

allows data-driven flexibilities. First, for the m-th expert, a prior gating weight w_m^{p} is computed by the Euclidean distance d_m between its cluster centroid and the semantic embedding of the input x:

$$w_m^{\mathrm{p}} = \frac{\exp(-d_m)}{\sum_{m'=1}^{n} \exp(-d_{m'})}. \quad (5)$$

Note that for experts associated with multiple clusters with single disease types, the smallest distance to their centroids is selected.

Second, the embedding of x is concatenated with the encoding of all experts and processed by a simple feedforward neural network. This produces the data-driven gating weights $\boldsymbol{w}^{\mathrm{d}} = (w_1^{\mathrm{d}}, \ldots, w_m^{\mathrm{d}})$. The final gating weights are computed as the average of the two:

$$w_m = \frac{1}{2}(w_m^{\mathrm{p}} + w_m^{\mathrm{d}}). \quad (6)$$

3.4 Model Training

The experts and gating network of SGMoE are jointly trained with the training reports and their disease labels. To improve the learning of the gating network, in additional to the commonly used cross-entropy classification loss $\mathcal{L}_{\mathrm{c}}$, we also develop a gating loss $\mathcal{L}_{\mathrm{g}}$ that guides the gating based on the prior knowledge of expert assignment. Specifically, for an input sample x with a true label $y \in \{a_1, \ldots, a_D\}$, the gating loss $\mathcal{L}_{\mathrm{g}}$ is designed to encourage greater gating weights w_m of experts E_m that are assigned the class y:

$$\mathcal{L}_{\mathrm{g}} = -\log \sum_{m=1}^{M} \mathbb{I}(y \in \mathcal{C}_m) \cdot w_m, \quad (7)$$

where $\mathbb{I}(y \in \mathcal{C}_m)$ is an indicator function that equals one if $y \in \mathcal{C}_m$ and zero otherwise. The total loss function $\mathcal{L}$ combines $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{g}}$ with a weighting factor λ:

$$\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \lambda \mathcal{L}_{\mathrm{g}}. \quad (8)$$

λ is empirically set to 0.1 so that the primary focus remains on classification while encouraging desired expert activation.

3.5 Implementation Details

The experts in SGMoE use the RoBERTa backbone (Cui et al., 2020, 2019), but other architecture may also be used. The pretrained RoBERTa extracts the semantic embedding of reports for clustering. It is then fine-tuned with the parameter efficient fine-tuning method LoRA (Hu et al., 2021) based on training data with a rank of 16 and an alpha of 32, where task-specific classification layers are attached and also learned. Different experts are initialized with unique random seeds. The gating network employs three fully connected layers. The AdamW optimizer is used for training with a learning rate of 2×10^{-5}, a batch size of 32, and 100 epochs.

Disease Type	Count	Disease Type	Count
Lymphoma	579	Metastases	1,250
Glioblastoma	1,231	Circumscribed glioma	534
Acoustic neuroma	778	Cavernous angioma	627
Neuronal tumor	697	Ependymoma	372
Medulloblastoma	308	Germinoma	622
Meningioma	1,250	Craniopharyngioma	1,096
Pituitary adenoma	1,250	Epidermoid cyst	541
Chordoma	361	Hemangioblastoma	368
Total	**11,864**	**-**	

Table 1: Composition of the In-house Dataset

4 Experiments

4.1 Dataset and Task Description

We performed experiments on two datasets. The first one was an in-house dataset consisting of 11,864 radiology reports of brain MRI, covering 16 different types of brain tumors, as summarized in Table 1. The ground truth classification was obtained with pathological analysis. The second dataset was the public CT-RATE dataset, containing radiology reports of chest CT images annotated with 18 different clinical manifestations. Each combination of the manifestations was considered a class. Only combinations with at least three distinct clinical manifestations were kept, which led to 40 classes with 1,862 samples. The detailed composition of these different classes is summarized in Table 2, where for convenience each manifestation combination is given a class index.

Both datasets were divided into training, validation, and test sets with a 8:1:1 ratio for each class. The number of experts in SGMoE determined by the semantic clustering results was seven for the in-house dataset and 12 for the CT-RATE dataset. The detailed expert assignments are shown in Tables 3 and 4 for the in-house and public datasets, respectively.

4.2 Disease Classification Results

SGMoE was compared with several text classification models based on popular text encoders,

Class (Manifestation Combination)	Index	Count	Class (Manifestation Combination)	Index	Count
AWC/AT/CAWC	1	27	AT/LN/LO/MM	21	21
AWC/AT/CAWC/EM/HH/LN/LA	2	20	AT/LN/LA/PFS	22	26
AWC/AT/CAWC/EM/LN/PFS	3	40	AT/LN/PFS	23	100
AWC/AT/CAWC/LN	4	52	AT/MM/PE	24	32
AWC/AT/CAWC/LN/LO	5	22	BR/EM/LN	25	38
AWC/AT/CAWC/LN/PFS	6	22	BR/LN/LO	26	34
AWC/AT/CAWC/LO	7	35	BR/LN/PBT	27	64
AWC/CO/CAWC/HH/LO	8	20	BR/LN/PFS	28	88
AWC/CO/CAWC/LO	9	59	CO/HH/LO	29	55
AWC/CAWC/EM/HH/LN/PFS	10	24	CO/LN/LO/PFS	30	58
AWC/CAWC/EM/LN	11	55	CO/LO/MM/PE	31	20
AWC/CAWC/EM/LN/LA/PFS	12	43	EM/LN/LA	32	60
AWC/CAWC/EM/LN/PBT/PFS	13	22	EM/LN/LA/PFS	33	46
AWC/CAWC/EM/PFS	14	32	HH/LN/LO	34	43
AWC/CAWC/LN	15	100	HH/LN/LA	35	44
AWC/CAWC/LN/LA/PFS	16	36	HH/LN/PFS	36	100
AWC/LN/PFS	17	67	HH/LO/LA	37	34
AT/CO/LN	18	20	LN/LO/PFS	38	100
AT/CO/LO	19	96	LN/LA/MM	39	33
AT/EM/LN/PFS	20	33	LN/MA/PFS	40	41
Total	**-**	**1,862**		**-**	

Table 2: Composition of the public CT-RATE dataset with manifestation combinations.AWC=Arterial Wall Calcification, AT=Atelectasis, CAWC=Coronary Artery Wall Calcification, EM=Emphysema, HH=Hiatal Hernia, LN=Lung Nodule, LA=Lymphadenopathy, PFS=Pulmonary Fibrotic Sequela, CO=Consolidation, LO=Lung Opacity, MM=Medical Material, PE=Pleural Effusion, BR=Bronchiectasis, PBT=Peribronchial Thickening, MA=Mosaic Attenuation Pattern.

Expert	Class Assignment
Expert 1	Lymphoma | Metastases | Glioblastoma | Ependymoma
Expert 2	Cavernous angioma | Epidermoid cyst
Expert 3	Circumscribed glioma | Germinoma | Craniopharyngioma
Expert 4	Circumscribed glioma | Neuronal tumor | Ependymoma
Expert 5	Circumscribed glioma | Hemangioblastoma
Expert 6	Circumscribed glioma | Ependymoma | Medulloblastoma
Expert 7	Lymphoma | Metastases | Acoustic neuroma | Germinoma | Meningioma | Pituitary adenoma | Chordoma

Table 3: Expert Assignment for In-house Dataset

Expert	Class Assignment
Expert 1	1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16
Expert 2	1 | 4 | 19 | 23 | 25 | 27 | 28 | 29 | 30 | 33 | 39
Expert 3	21 | 24 | 31 | 39
Expert 4	22 | 32 | 33 | 35 | 37 | 39
Expert 5	5 | 18 | 19 | 20 | 22 | 23 | 26 | 27 | 28 | 30 | 34 | 38 | 40
Expert 6	8 | 9 | 19 | 29 | 30 | 31
Expert 7	10 | 29 | 34 | 35 | 36 | 37
Expert 8	13 | 25 | 27
Expert 9	2 | 3 | 11 | 12 | 14 | 32
Expert 10	6 | 10 | 12 | 14 | 17
Expert 11	2 | 12 | 17
Expert 12	1 | 4 | 6 | 13 | 15 | 17 | 18 | 19 | 20 | 22 | 23 | 25 | 26 | 27 | 28 | 29 | 30 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40

Table 4: Expert assignment for CT-RATE dataset. Indexes in Table 2 are used to indicate the corresponding categories.

including LSTM (Hochreiter and Schmidhuber, 1997), BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), and RoBERTa (Cui et al., 2020, 2019). LSTM represents a traditional sequential model commonly used for text classification, whereas BioBERT and ClinicalBERT are Transformer-based models pretrained specifically on biomedical and clinical text, respectively. RoBERTa is a robustly optimized general-purpose language model, often serving as a strong baseline for NLP tasks. The same task-specific classification layers of SGMoE were attached to these encoders for disease classification based on radiology reports, and the competing models were trained on the same training data and under the same settings with LoRA (except for the small-capacity LSTM with full-parameter fine-tuning) as SGMoE to ensure fair comparison.

The classification performance was quantitatively measured by the average classification accuracy (ACC) and average F1-score. The results are shown in Table 5, where SGMoE shows clear

Model	In-house		CT-RATE	
	ACC (%)	F1 (%)	ACC (%)	F1 (%)
LSTM	54.6	55.0	72.7	73.6
BioBERT	53.3	51.1	69.0	65.4
ClinicalBERT	62.4	61.4	73.8	69.0
RoBERTa	70.1	69.3	88.2	87.7
SGMoE	**72.0**	**71.4**	**92.5**	**92.5**

Table 5: Classification performance comparison of SGMoE with baseline models

advantages over the competing methods. For the in-house dataset, SGMoE achieves an average ACC of 72.0% and an average F1-score of 71.4%, outperforming the best competing method RoBERTa by 1.9 and 2.1 percentage points, respectively; compared with other competing methods, the improvement in ACC and F1-score is close to or even greater than 10 percentage points. For the public CT-RATE dataset, SGMoE achieves an average ACC of 92.5% and an average F1-score of 92.5%, outperforming the best competing method RoBERTa by 4.3 and 4.8 percentage points, respectively; compared with other competing methods, the improvement in ACC and F1-score is close to or even greater than 20 percentage points.

As SGMoE also uses the RoBERTa backbone, the comparison between SGMoE and RoBERTa directly indicates the benefit of expert specialization guided by semantic clustering, where the decomposition of the multi-class disease classification problem has led to improve classification performance.

4.3 Comparison with Medical Expert

In addition, we compared the performance of SGMoE with the classification made by a medical expert based on the report. As the radiologist evaluation was manual, only a subset of the test set was selected from the in-house dataset. Specifically, 96 cases were randomly selected (six samples for each disease type). The results are shown in Table 6. SGMoE achieved better accuracy than the radiologists with an improvement of 14.5% (63.5% vs. 49.0%). The improvement indicates that SGMoE not only is superior to other automated methods but also can surpass medical experts when addressing the multi-class classification problem.

4.4 Ablation Study

To verify the contributions of the key components in SGMoE, we performed a series of ablation experiments on the in-house dataset. Each ablation setting isolated the effect of a specific component. The results are summarized in Table 7 and explained below.

Evaluator	Acc (%)
SGMoE	63.5
Medical expert	49.0

Table 6: Classification performance comparison between SGMoE and a medical expert

Variant	ACC (%)	F1 (%)
SGMoE (Full)	72.0	71.4
Without prior gating weight	71.6	70.9
Without gating loss	70.3	69.6
Without class routing	71.4	70.8
Random class allocation	69.9	69.4

Table 7: Ablation study on the in-house dataset

Prior Gating Weight First, we removed the semantic prior gating weights from the gating network. This forced the gating mechanism to rely solely on sample-specific features to determine expert activations without guidance from the semantic clustering results. The removal led to a decrease in performance, with ACC dropping from 72.0% to 71.6% and F1-score from 71.4% to 70.9%. The performance decrease indicates that the gating prior is beneficial to the activation of appropriate experts.

Gating Loss Next, we removed the gating loss and trained the whole network with the classification loss only. This adjustment removed the explicit constraint that aligns the gating weight with the predefined expert assignments. The ACC dropped to 70.3% and F1-Score dropped to 69.6%. This result demonstrates that the gating loss is important for the gating network to learn how to activate relevant experts, thereby improving the overall performance.

Class Routing We also removed the class routing mechanism, allowing each expert to predict all 16 classes. This change reduced the ability of the experts to specialize in semantically meaningful clusters, leading to decreased ACC (71.4%) and F1-score (70.8%). The performance decrease highlights the importance of class routing in ensuring that each expert focuses on a specific subset of the problem space, which reduces the complexity of the classification task.

Expert Assignment SGMoE uses semantic clustering to guide expert assignment. To show the benefit of this assignment, we replaced the semantics-guided assignment with random assignment, where we generated random disease type groups while retaining the number of classes for each expert and ensuring all disease types were covered.

This resulted in the largest performance degradation among all settings, with ACC dropping to 69.9% and F1-score dropping to 69.4%. The performance is even worse than that of RoBERTa. These results emphasize the critical role of semantic clustering in defining distinct and meaningful subspaces for the experts, enabling them to specialize and collectively improve the classification performance.

4.5 Cluster Analysis

As we assume that semantic clustering allows multimodal distributions to be decomposed into simpler distributions, to verify the assumption, we performed cluster analysis on the in-house dataset. Specifically, we selected the circumscribed glioma that had multiple modes in the space of semantic embeddings and used t-SNE (Van der Maaten and Hinton, 2008) to project the high-dimensional embeddings of circumscribed glioma samples into a two-dimensional space for visualization. The visualization result is shown in Fig. 2. The data points belonging to different clusters are colored differently, and they are distributed across four clusters (Cluster 3, 4, 5, and 6). The samples scatter in the embedding space but in each cluster the samples are relatively compact.

The visualization shows that our assumption is valid and it is reasonable to assign different experts to handle different distribution modes.

5 Conclusion

In this paper, we have proposed the SGMoE framework for multi-class disease classification based on radiology reports. SGMoE comprises multiple experts that handle different subproblems decomposed from the original challenging problem. In particular, semantic clustering is applied to guide the expert assignment. With the given assignment, SGMoE further develops a gating mechanism that takes the prior clustering knowledge into consideration and a gating loss that facilitates the learning of gating. Experimental results on an in-house dataset and a public dataset show that SGMoE consistently outperforms competing text classification models.

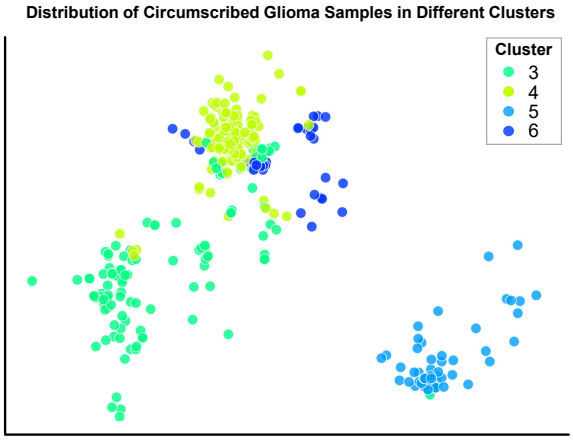

Figure 2: Visualization of sample distributions for circumscribed glioma with a t-SNE plot. Different modes can be assigned to different clusters obtained with semantic embeddings.

Limitations

There are limitations of SGMoE. First, it handles scenarios with complex class distributions, and for simple classification problems its gains may become minimal. Second, in this work the medical expert compared with SGMoE was a junior radiologist. It may be interesting to further incorporate senior radiologists for comparison in future work and evaluate how SGMoE compares with more experienced experts.

Ethical statement

All the experiments strictly follow the ACL Code of Ethics. The in-house dataset was accessed with a formal data-sharing agreement with xxx hospital. The study was in accordance with the Declaration of Helsinki and approved by each center and their local ethics committees. All participants were informed of the details of this study and signed informed consent forms before the interviews; the data usage was strictly limited to this research. The public dataset was cited in Section 1, adhering to its original license terms. The dataset was retrospectively collected with approval from xxx Institutional Review Board. No direct interaction with patients occurred in this study. The publicly available code for the competing methods is cited in Section 4.2, which follows their original license. Our implementation code will be open-sourced upon acceptance, excluding in-house data preprocessing scripts due to confidentiality agreements.

References

Mouhamed Laid Abimouloud, Khaled Bensid, Mohamed Elleuch, Oussama Aiadi, and Monji Kherallah. 2024. Vision transformer-convolution for breast cancer classification using mammography images: A comparative study. *International Journal of Hybrid Intelligent Systems*, 20(2):67–83.

Zeynettin Akkus, Issa Ali, Jiri Sedlar, Timothy L. Kline, Jay P. Agrawal, Ian F. Parney, Caterina Giannini, and Bradley J. Erickson. 2016. Predicting 1p19q Chromosomal Deletion of Low-Grade Gliomas from MR Images using Deep Learning. *Preprint*, arXiv:1611.06939.

Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. 2017. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162.

Matthew C Chen, Robyn L Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E Chapman, David B Larson, Curtis P Langlotz, Timothy J Amrhein, and Matthew P Lungren. 2018. Deep learning to classify radiology free-text reports. *Radiology*, 286(3):845–852.

Mengyao Chen, Meng Zhang, Lijuan Yin, Lu Ma, Renxing Ding, Tao Zheng, Qiang Yue, Su Lui, and Huaiqiang Sun. 2024. Medical image foundation models in assisting diagnosis of brain tumors: a pilot study. *European Radiology*, pages 1–13.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Yoon Seong Choi, Sohi Bae, Jong Hee Chang, Seok-Gu Kang, Se Hoon Kim, Jinna Kim, Tyler Hyungtaek Rim, Seung Hong Choi, Rajan Jain, and Seung-Koo Lee. 2021. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro-oncology*, 23(2):304–313.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xin Gao, Meihui Zhang, Longfei Chen, Jun Qiu, Shanbo Zhao, Junjie Li, Tiantian Hua, Ying Jin, Zhiqiang Wu, Haotian Hou, and 1 others. 2024. Simple Words over Rich Imaging: Accurate Brain Disease Classification via Language Model Analysis of Radiological Reports. *medRxiv*, pages 2024–11.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. 2024a. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. *Preprint*, arXiv:2403.17834.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024b. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. *Preprint*, arXiv:2403.06801.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, Chinmay Prabhakar, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. 2024c. GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes. *Preprint*, arXiv:2305.16037.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Md Zabirul Islam, Md Milon Islam, and Amanullah Asraf. 2020. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in medicine unlocked*, 20:100412.

Anil K Jain and Richard C Dubes. 1988. *Algorithms for clustering data*. Prentice-Hall, Inc.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Chulho Kim, Vivienne Zhu, Jihad Obeid, and Leslie Lenert. 2019. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroketang2020automated. *PloS one*, 14(2):e0212778.

Panagiotis Korfiatis, Timothy L Kline, Daniel H Lachance, Ian F Parney, Jan C Buckner, and Bradley J Erickson. 2017. Residual deep convolutional neural network predicts MGMT methylation status. *Journal of digital imaging*, 30:622–628.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Sicen Liu, Xiaolong Wang, Yang Xiang, Hui Xu, Hui Wang, and Buzhou Tang. 2022. Multi-channel fusion LSTM for medical event prediction using EHRs. *Journal of Biomedical Informatics*, 127:104011.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Milad Mirbabaie, Stefan Stieglitz, and Nicholas RJ Frick. 2021. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health and Technology*, 11(4):693–731.

Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, Chang Min Park, and Jong Chul Ye. 2022. Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nature communications*, 13(1):3848.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, and 1 others. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(2):53–65.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. 2021. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795.

Ahmad Waleed Salehi, Preety Baglat, Brij Bhushan Sharma, Gaurav Gupta, and Ankita Upadhya. 2020. A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pages 156–161. IEEE.

Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and 1 others. 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*, 3(1):70.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Wei Wang, Ran Jiang, Ning Cui, Qian Li, Feng Yuan, and Zhifeng Xiao. 2022. Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. *Frontiers in Pharmacology*, 13:929755.

Anthony B Wolbarst and William R Hendee. 2006. Evolving and experimental technologies in medical imaging. *Radiology*, 238(1):16–39.

Xiaowei Xu, Qianjun Jia, Haiyun Yuan, Hailong Qiu, Yuhao Dong, Wen Xie, Zeyang Yao, Jiawei Zhang, Zhiqaing Nie, Xiaomeng Li, and 1 others. 2023. A clinically applicable AI system for diagnosis of congenital heart diseases based on computed tomography images. *Medical Image Analysis*, 90:102953.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. RadBERT: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7):1235–1270.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. 2019. Attention residual learning for skin lesion classification. *IEEE transactions on medical imaging*, 38(9):2092–2103.