applied sciences

MDPI

Article

MSA K-BERT: A Method for Medical Text Intent Classification

Yujia Yuan and Guan Xi *

School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, 100 Waihuanxi Road, Guangzhou 510006, China; 3222008468@mail2.gdut.edu.cn
* Correspondence: xiguan0704@gdut.edu.cn

Abstract: Improving medical text intent classification accuracy can assist the medical field in achieving more precise diagnoses. However, existing methods suffer from problems such as low accuracy and a lack of knowledge supplementation. To address these challenges, this paper proposes MSA K-BERT, a knowledge-enhanced bidirectional encoder representation model that integrates a multi-scale attention (MSA) mechanism to enhance prediction performance while solving critical issues like heterogeneity of embedding spaces and knowledge noise. We systematically validate the reliability of this model on medical text intent classification datasets and compare it with various deep learning models. The research results indicate that MSA K-BERT makes the following key contributions: First, it introduces a knowledge-supported language representation model compatible with BERT, enhancing language representations through the refined injection of knowledge graphs. Second, it adopts a multi-scale attention mechanism to reinforce different feature layers, significantly improving the model's accuracy and interpretability. Especially in the IMCS-21 dataset, MSA K-BERT achieves precision, recall, and F_1 scores of 0.826, 0.794, and 0.810, respectively, all exceeding the current mainstream methods.

Keywords: medical text intent classification; natural language processing; knowledge-enhanced BERT; IMCS-21; multi-scale attention

Academic Editor: Douglas O'Shaughnessy

Received: 8 May 2025
Revised: 9 June 2025
Accepted: 15 June 2025
Published: 17 June 2025

Citation: Yuan, Y.; Xi, G. MSA K-BERT: A Method for Medical Text Intent Classification. *Appl. Sci.* **2025**, *15*, 6834. https://doi.org/10.3390/app15126834

1. Introduction

Medical text intent [1] refers to identifying the intended purpose of a paragraph from medical field texts, including patient questions, doctor responses, and medical records in natural language processing (NLP), providing an understanding basis for medical NLP systems. Due to the fact that medical texts contain many obscure medical professional terms and usually do not follow natural grammar [2], medical text intent classification is confronted with problems such as poor model generalization ability and interpretability, low quality and quantity of data annotation, and the need for updates in the construction of the classification system.

Medical text intent classification methods have evolved from rule-based template matching to data-driven approaches. Early rule-based methods [3] are being gradually replaced by machine learning methods [4–6]. Modern models no longer rely on manually defined rule templates; instead, they automatically learn semantic patterns and decision boundaries from labeled data through algorithms. In recent years, deep learning methods [7,8] become mainstream due to their ability to perform automatic feature extraction.

In deep learning approaches, convolutional neural networks (CNNs) [9] and Bidirectional Encoder Representations from Transformers (BERTs) [10] currently represent state-of-the-art methods for many medical text intent classification tasks.

Single CNN models [11] such as Text attention CNN (Text-CNN) [12,13] and Dynamic CNN (DCNN) [14] typically rely on convolutional operations to extract local features, though they may exhibit limitations in modeling long-range dependencies and sequence positional information. Hybrid CNN [15] architectures that integrate Recurrent Neural Network (RNN) [9] or Attention mechanisms [16], such as CNN-RNN [17] and CNN-Attention [18] models, effectively address these limitations. While these approaches demonstrate strong performance on biomedical datasets, including PubMed 200 K RCT [19] and BioASQ [20], they continue to present several research challenges regarding the optimal configuration of convolution kernels, handling of imbalanced data distributions, and effective incorporation of external knowledge and domain expertise.

BERT-based models for medical text intent classification [21,22] have emerged as a new mainstream approach due to their pre-trained semantic representation capabilities. Single BERT models [22] (e.g., BERT-base [21], BERT-large [23]) typically achieve a strong baseline performance through direct fine-tuning, while hybrid BERT [24] architectures (e.g., BERT-CNN [25], BERT-BiLSTM [26]) further combine the advantages of local feature extraction and contextual modeling. In tasks such as the classification of randomized controlled trial titles, abstracts, biomedical question answering, and medical literature retrieval, BERT-based models generally demonstrate superior performance compared to traditional deep learning models. However, they still face limitations in optimizing pre-training strategies and processing multimodal data. Particularly when incorporating external knowledge, challenges like heterogeneous embedding space (HES) [27] and knowledge noise (KN) [28] remain critical bottlenecks affecting model performance.

Moreover, although Large Language Models (LLMs) [29] like GPT [30] and Large Language Model Meta AI (LLaMA) [29] achieve impressive results in tackling various complex tasks, they perform poorly in domain-specific classification tasks due to limitations such as resource constraints and explainability. Mullick et al. [31] study the performance of various models on medical intent classification tasks. The results show that models specifically pre-trained for the medical domain, such as RoBERTa [32] and PubMedBERT [31], significantly outperform general-purpose large language models like ChatGPT [30] and LLaMA-2 [33] across multiple datasets. Specifically, on the CMID dataset [1], RoBERTa achieves an accuracy of 72.88%, while ChatGPT achieves only 42.36%. Therefore, compared to relying on general-purpose LLMs, choosing specialized models is a more reliable option for handling domain-specific tasks that require high precision.

To address these challenges, we propose MSA K-BERT, which is a knowledge-enhanced bidirectional encoder representation model. It integrates the multi-scale attention mechanism (MSA) [34] to enhance the prediction performance by selectively focusing on the text content. This model is applicable to the medical text intent classification task. Due to its parameter consistency, MSA K-BERT is compatible with any pre-trained BERT model. Furthermore, it can effortlessly inject domain-specific knowledge related to the medical field into the model.

In light of the above, this paper presents the following key contributions:

1. A knowledge-supported Language Representation (LR) model compatible with BERT is proposed in this paper, namely MSA K-BERT, which can combine knowledge in the medical field and alleviate the problems of HES and KN.
2. Through a fine-grained injection of a Knowledge Graph (KG), the performance of MSA K-BERT in the medical text intent classification task is superior to that of mainstream models such as BERT.
3. Using the multi-scale attention mechanism to enhance the feature layers at different stages and selectively assign different weights to the text content, MSA K-BERT can enhance the text content with the attention mechanism, which not only makes the results more accurate but also makes them more interpretable.

2. Background

2.1. Knowledge Graph

A Knowledge Graph (KG) [35] typically serves as a graph-structured framework for organizing domain knowledge, with entity relationships commonly represented in ⟨head, relation, tail⟩ triple form. In the medical domain, KG is widely utilized to integrate multi-source heterogeneous data and construct structured medical knowledge systems. It typically provides semantic foundations and explainability support for tasks such as medical text intent classification. For instance, when a patient describes symptoms like "fever and sore throat", a KG-based model can leverage triple relationships such as ⟨fever, common_symptom_of, common_cold⟩ and ⟨sore_throat, common_symptom_of, common_cold⟩ to infer the patient's potential inquiry about cold-related medication advice.

2.2. Knowledge Noise

In medical text intent classification tasks, such as identifying patient consultation intents, parsing clinical instructions, or inferring query intents for medical terminology, the performance of models tends to be influenced by Knowledge Noise (KN) [28]. KN refers to interference factors in the text coupled with domain knowledge, including variations and abbreviations of medical terms (e.g., "myocardial infarction" and "MI"), non-standard expressions (e.g., a patient describing "chest discomfort" to mean chest pain), and contextual ambiguities. Such noise is not a random error but rather interference derived from the medical knowledge system. It may distort semantic representations, potentially blurring intent boundaries in models, and could introduce safety risks through the misclassification of clinical instructions. Research indicates that KN may lead to significant accuracy degradation in intent classification. Studies such as those by Sengupta et al. [36], which evaluated seven common noise types, suggest that even state-of-the-art BERT-based intent classifiers typically experience an average accuracy reduction of approximately 13.5% under noisy conditions. Consequently, mitigating KN interference in medical text intent classification constitutes a critical prerequisite for enhancing model performance.

2.3. Heterogeneity of Embedding Spaces

Heterogeneity of Embedding Spaces (HES) [27,37] refers to inconsistencies in the embedded representations of words, entities, or other linguistic elements, which may arise from variations in contextual, syntactic, or semantic attributes. Such divergence often leads to incompatibility in vector space representations, making it challenging for models to effectively integrate and utilize these embeddings for text intent classification tasks. In medical texts, the frequent presence of abbreviations, domain-specific terms, and informal expressions may further exacerbate semantic discrepancies. To address challenges posed by HES, researchers have developed various optimization techniques. For instance, Khalid et al. [38] propose a label-supervised contrastive learning approach that enhances inter-class discrimination by introducing label anchors in both Euclidean and hyperbolic embedding spaces, demonstrating improved performance on imbalanced text classification tasks. However, how to enhance the model's robustness to the HES while maintaining its generalization ability remains an important direction in current research.

3. Related Work

3.1. Text Classification

Early text classification methods are primarily based on machine learning techniques, which focus on enabling computer systems to learn from data and continuously promote their performance without the need for explicit programming. This technology makes computers learn from experience and make informed decisions or predictions on new data.

Mitra et al. [39] design a Least Squares Support Vector Machine (LS-SVM) for text classification of noisy document titles. The method achieves a classification accuracy exceeding 99% when evaluated on a corpus of 91,229 words from the Penrose Library catalog at the University of Denver. Isa et al. [40] implement a hybrid enhancement method using Naive Bayes and SVM, which improves the average classification accuracy by 1.05% compared to the standard SVM and by 4.41% compared to the Bayesian classifier. To effectively extract complex features from unbalanced text data, Wu et al. [41] propose an ensemble method based on Random Forest (RF), called ForesTexter. This model outperforms Random Forest in fitting imbalanced data, and its AUC score is 3.27% higher than the corresponding score of RF.

Machine learning algorithms perform well when dealing with small-scale data [42], mainly because of their low computational resource requirements and fast model training. However, traditional machine learning methods have some limitations when dealing with large-scale, high-dimensional sparse data. Therefore, to solve the above-mentioned problems, the emergence of deep learning techniques [8,43] has brought significant breakthroughs and innovations in the text classification field.

The introduction and improvement of deep learning models have achieved significant performance enhancements in the text classification task. Huang et al. [13], considering the significant impact of noisy background information on extracting real text, propose a new text attention convolutional neural network, namely Text-CNN, which emphasizes extracting text features from image components and distinguishing effective text regions. Compared with the traditional CNN model, this model has increased the average accuracy of text recognition and classification to 91%. To effectively handle the high dimensionality, sparsity, and complex semantics of text data, Liu et al. [44] propose the AC-BILSTM model, which is a bidirectional long short-term memory network with an attention mechanism and convolutional layers. This model is evaluated on seven comprehensive labeled datasets. In comparison with the previous state-of-the-art text classification methods, this model demonstrates outstanding classification accuracy and robustness. In a multidimensional sentiment text classification analysis, Jin et al. [45] propose a tree-structured CNN-LSTM model, composed of region-based CNN and LSTM, to predict sentiment information ratings. By comparing with other LSTM and CNN models, the results indicate that the Regional CNN-LSTM (Tree) model, which combines region division with tree depth structure information, exhibits outstanding predictive performance. The model achieves a Mean Squared Error (MSE) of 0.94, and the observed performance differences are statistically significant ($p < 0.05$).

It can be seen that neural network-based deep learning models [46,47] have high accuracy and performance when dealing with large-scale or high-dimensional data. However, neural network models usually require substantial labeled data, computational resources, and a long training time when performing large-scale training tasks [48,49]. In terms of training effectiveness, neural network models still lack significant generalization ability when applied to different domains or tasks. Over the past few years, pre-training models such as BERT [10] have been proposed and widely used as training strategies for deep learning models. By performing self-supervised learning on large-scale unlabeled data, these models are able to capture more generalized semantic features, thus improving their performance.

3.2. Pretrained Language Models

Thanks to the BERT model proposed by Devlin et al. [10], pre-training and fine-tuning have become common methods for NLP tasks. Due to BERT's self-attention mechanism that enables it to capture contextual information from the whole sequence, BERT shows

superior performance on several NLP tasks, especially question–answer and linguistic reasoning, by considering both the context on the left side and the context on the right side of a word. Unlike traditional methods, BERT employs a Masked Language Model (MLM) as its pre-training target. MLM aims to randomly mask specific words within a sentence by substituting them with the [mask] token. Its subsequent objective is to infer the original masked words by leveraging contextual information from both preceding and following words. BERT handles downstream classification tasks by using the first token of each sequence, known as the special classification token [CLS]. In BERT, the [CLS] token in the final hidden layer aggregates the representation of a sentence or a pair of sentences. Although BERT performs exceptionally well and is applicable to various NLP tasks, it also has some shortcomings.

To tackle the issue of BERT's neglect of knowledge integration in language understanding, Sun et al. [50] propose Enhanced Language Representation with Informative Entities (ERNIE) for learning language representations. The language representation is enhanced by a knowledge masking strategy. Liu et al. [32] take advantage of using a large and extended corpus. They develop a powerful BERT pre-training method called a Robustly optimized BERT approach (RoBERTa). In addition, the loss function and scalability of BERT are also key concerns for researchers. Clark et al. [51] propose Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), a model that introduces a binary classification loss function. ELECTRA uses generators and discriminators to speed up the learning process. Lan et al. [52] design A Lite BERT (ALBERT) architecture, which introduces a self-supervised loss for sentence order prediction based on BERT, along with a parameter sharing mechanism and factorized embedding parameterization to improve model efficiency and scalability, address memory constraints, and reduce communication overhead. ERNIE, RoBERTa, ELECTRA, and ALBERT lead in most NLP benchmarks, including SQuAD [53] and GLUE [54].

3.3. Knowledge Noise and Heterogeneity of Embedding Spaces

The presence of KN, such as spelling errors, abbreviations, and nonstandard expressions, may affect models' semantic judgment and reduce their classification performance. Moradi et al. [55] design 16 noise injection methods to simulate real-world clinical scenarios, testing three models—ClinicalBERT, ClinicalXLNet, and ClinicalELMo—on four NLP tasks. The results suggest that noise tends to cause varying degrees of performance decline across all models. Naseem et al. [28] point out that KG-based language models demonstrate strong performance in biomedical NLP tasks, though they may face challenges such as KN and neglecting contextual relationships. To address this, the authors propose two novel approaches: combining KG language models with nearest-neighbor models or integrating them with Graph Neural Networks (GNNs). Experimental results indicate that these methods can lead to notable performance improvements in relation to extraction and classification tasks. HES refers to the embedding representations of words and entities in the text, which are often different, leading to incompatibility in their vector spaces. This mismatch can affect the accuracy of classification models. Si et al. [56] compare traditional word embeddings with contextual embeddings in clinical NLP tasks, observing that contextual embeddings tend to capture semantic information more effectively and show potential for improving model performance. Zhang et al. [57] observe that pretrained BERT models encode biases related to gender, race, and language in clinical texts, which may lead to reduced performance for minority groups in downstream tasks. Therefore, HES affects the fairness and accuracy of the model.

4. Methods

4.1. Model Architecture

We represent a sentence as $s = \{h_0, h_1, h_2, \ldots, h_n\}$, where "$n$" represents the sentence's length. In this work, English tokens are based on words, while Chinese tokens are based on characters. Every token h_i belongs to the vocabulary V, that is, $h_i \in V$. KG is denoted as K, and it is a collection of triples $\varepsilon = (h_i, r_j, t_k)$, where h_i and t_k are the names of entities, and $r_j \in V$ represents the relationship between them. All triples are part of the KG, that is, $\varepsilon \in K$.

As illustrated in Figure 1, the MSA K-BERT is composed of five modules: the Knowledge Layer, Embedding Layer, Seeing Layer, Mask Transformer, and Multi-Scale Attention Layer.

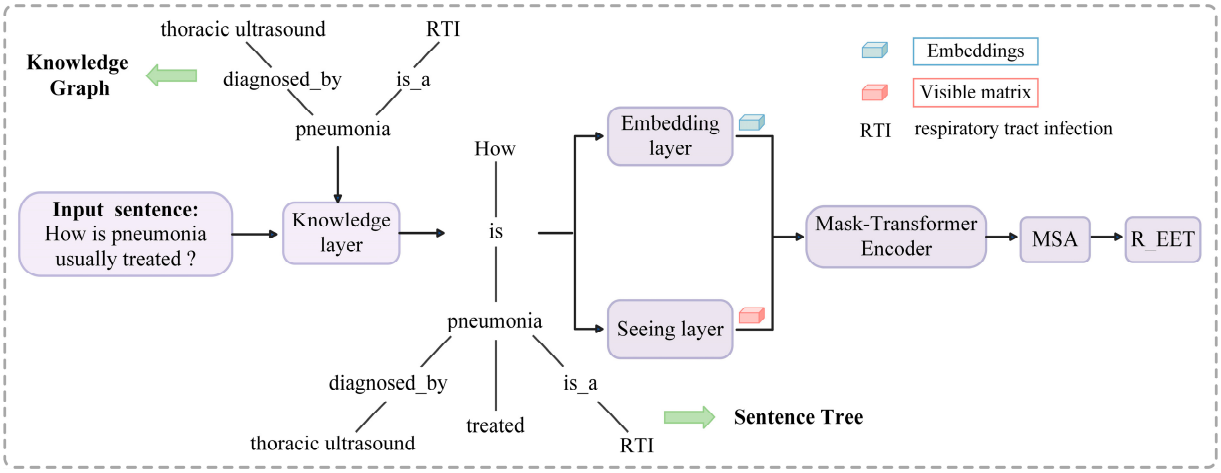

Figure 1. This is the model architecture diagram of MSA K-BERT. The abbreviation RTI stands for respiratory tract infection. Red cuboids denote the Visible matrix, whereas blue cuboids indicate the Embeddings. R_EET refers to the Request Existing Exam and Treatment category in the IMCS-21 database.

For medical intent documents, we divide the document into sentences, and each sentence is further split into words that contain domain-specific medical knowledge. The Knowledge Layer firstly introduces pertinent triples from the KG to convert the original sentence into a sentence tree enriched with knowledge. Subsequently, the sentence tree is input into the Embedding Layer and the Seeing Layer for further processing. After that, it is transformed into token-level embedding representations as well as a visibility matrix. The visibility range of each token is regulated through the visibility matrix, thereby preventing the original sentence semantics from being distorted when introducing external knowledge. During training, we adopt a full fine-tuning strategy, in which all parameters of the model—including the pretrained BERT backbone and the newly added modules such as the multi-scale attention layer—are updated jointly via backpropagation. This enables the entire architecture to better adapt to the medical intent classification task.

4.2. Knowledge Layer

The Knowledge Layer (KL) injects knowledge and generates sentence trees [58]. Specifically, when provided with an input sentence $s = \{h_0, h_1, h_2, \ldots, h_n\}$ and a KG K, the KL will construct a sentence tree $t = \{h_0, h_1, \ldots, h_i\,[(r_{i0}, t_{i0}), \ldots, (r_{ik}, t_{ik})], \ldots, h_n\}$. This process can be divided into two stages: Knowledge Query (K-Query) and Knowledge Injection (K-Inject). In the K-Query, all relevant entities are identified from the sentence s, and their corresponding triples are retrieved in the KG K. K-Query can be expressed as (1):

$$E = K - Query(s, K) \tag{1}$$

where $E = \{(h_i, r_{i0}, t_{i0}), \ldots, (h_i, r_{ik}, t_{ik})\}$ is the set of corresponding triples.

Subsequently, K-Inject combines the triples in E to their respective positions, injects the queried E into sentence s, and then constructs a sentence tree t. The construction of t is shown in Figure 2. The sentence tree is permitted to contain many branches, but its depth is constrained to 1, meaning that the entity names in the triples do not iteratively generate further branches. K-Inject is defined as (2):

$$t = K - Inject(s, E) \tag{2}$$

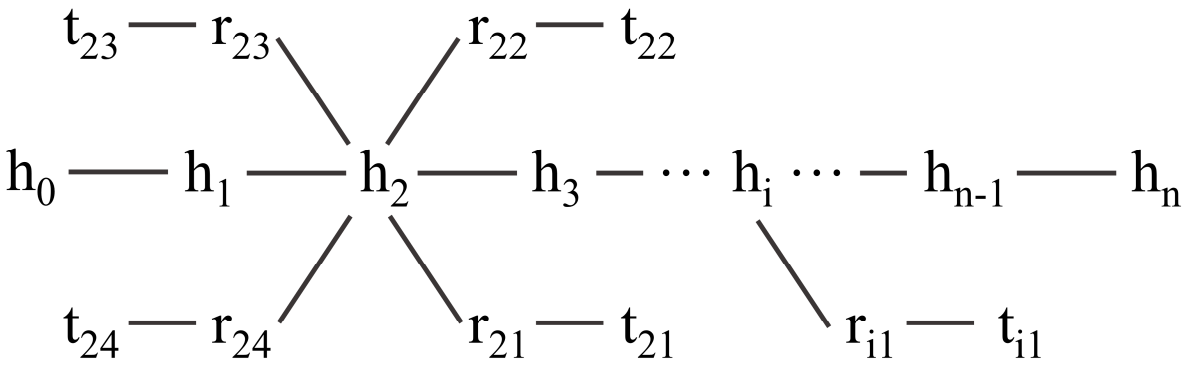

Figure 2. Structure of the sentence tree.

4.3. Embedding Layer

The Embedding Layer (EL) is a method that transforms text into numerical vectors, allowing computers to understand and process the text. The sentence tree is a graphical representation of the sentence structure and semantic relationships, composed of words and entities from the KG. The Mask-Transformer is a neural network model based on the self-attention mechanism, which can extract features and information from the input vectors. In this paper, MSA K-BERT is an LR model that combines BERT and KG. It leverages domain knowledge to enhance text understanding and generation capabilities. Unlike BERT, the input to MSA K-BERT is not a simple word sequence but a sentence tree that contains information from the KG. Therefore, K-BERT needs to address two issues in the EL: first, how to convert the sentence tree into a sequence of vectors and second, how to preserve the structural and semantic information within the sentence tree. To address these issues, MSA K-BERT introduces the concepts of soft position encoding and visibility matrices. These mechanisms help limit the interference of the KG on the sentence while preserving the original order and semantic relationships of the sentence.

4.4. Seeing Layer

An important distinction between MSA K-BERT and BERT is the presence of a seeing layer, which is also the reason why this method operates so efficiently. MSA K-BERT uses a sentence tree as its input, and the branches represent knowledge obtained from the KG. Nevertheless, the risk associated with knowledge fusion is that it may result in a change in the sentence's original meaning, which is the KN problem. For instance, in the sentence tree shown in Figure 3, the word [RTI] only modifies [pneumonia] and is unrelated to [treatment]. Therefore, the treatment of pneumonia should not be affected by the RTI. In addition, the [CLS] token employed for categorization should not be overlooked [pneumonia] for information about [RTI], as this may result in semantic distortion. To avoid this happening, MSA K-BERT uses a visibility matrix M to restrict the visible region of every token, ensuring that [CLS] and [RTI], [RTI] and [treated] are not visible to each other.

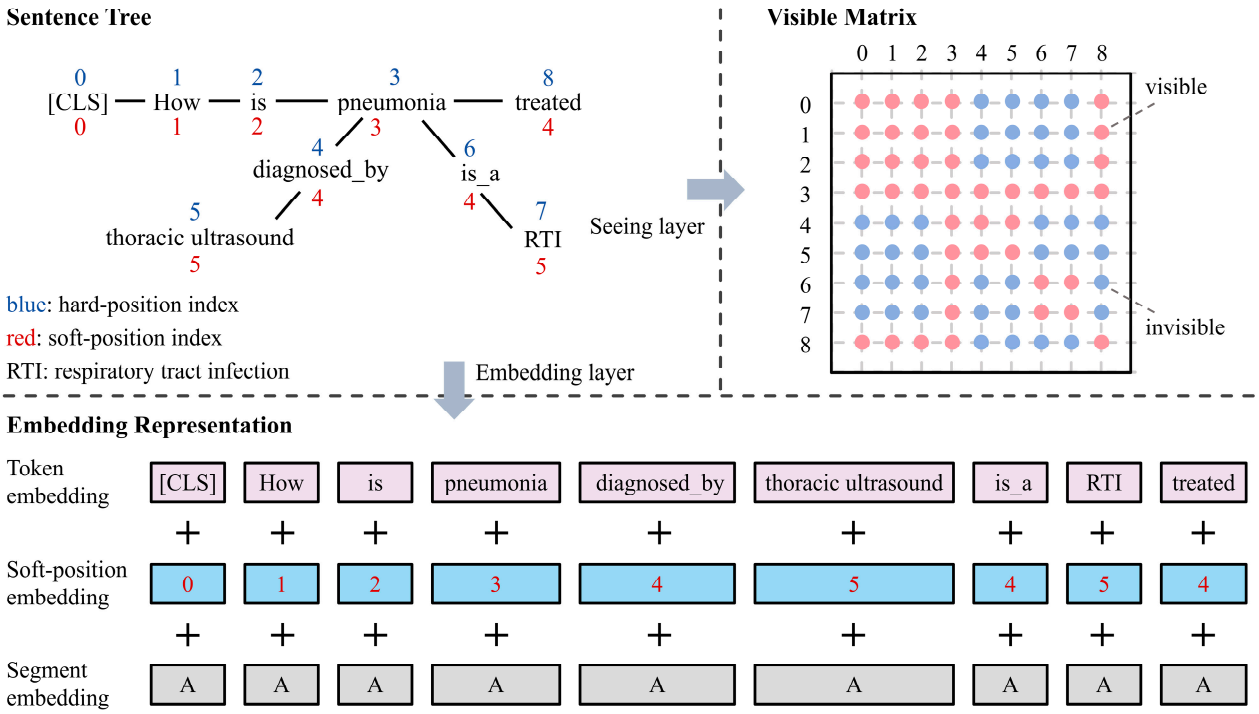

Figure 3. The procedure of transforming a sentence tree to an embedding representation and visible matrix. The hard-position index is represented by the blue number, and the soft-position index is represented by the red number in the sentence tree.

4.5. Mask-Transformer

The visibility matrix M contains information about the structure of the sentence tree to some extent. However, the transformer [59] encoder in BERT is unable to directly accept M as input. Therefore, we need to transform it into a Mask-Transformer, which restricts the self-attention area based on matrix M and consists of a stack of multiple Mask-Self-Attention blocks. Following the BERT architecture, L denotes the number of layers (i.e., the Mask-Self-Attention blocks), H denotes the hidden size, and A denotes the number of Mask-Self-Attention heads.

Mask-Self-Attention: In order to avoid semantic changes by leveraging the sentence construction information in M, a Mask-Self-Attention, which is an extension of the standard self-attention mechanism, is proposed. In form, it is represented as (3):

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v, \tag{3}$$

$$S^{i+1} = softmax\left(\frac{Q^{i+1}K^{i+1^T} + M}{\sqrt{d_k}}\right), \tag{4}$$

$$h^{i+1} = S^{i+1}V^{i+1}, \tag{5}$$

Among them, W_q, W_k, and W_v denote the trainable parameters of the model. h^i represents the hidden state output by the i-th Mask-Self-Attention block. d_k denotes the scaling factor. M denotes the visible matrix computed by the seeing layer. Intuitively speaking, if h_k is not visible to h_j, then M_{jk} will mask the attention score S_{jk}^{i+1} to 0, meaning that the hidden state of h_k does not contribute to h_j.

4.6. Multi-Scale Attention Layer

The multi-scale attention layer (MSA) is a method that applies weights to the input at different scales [60]. It can extract multi-level information and is suitable for processing text or images of different lengths or resolutions. One implementation method of the MSA is to use multiple convolution kernels of different sizes to convolve the input, then concatenate the convolution feature maps, calculate the attention weights through a fully connected layer or self-attention mechanism, and finally apply the attention weights to the input to obtain the multi-scale attention output.

For example, in the case of medical text intent classification, we assume the input is a matrix $X \in R^{n \times d}$, where n represents the text length. and d is the dimensionality of the word embeddings. We apply three convolutional kernels of different sizes, K_1, K_2, K_3 to convolve X, resulting in three feature maps of different scales: F_1, F_2, F_3, where each $F_i \in R^{n \times c}$ and c represents the number of output channels. Subsequently, we concatenate these three feature maps to obtain a concatenated feature map $F \in R^{n \times 3c}$. Next, a self-attention mechanism is employed to compute the attention weights $A \in R^{n \times n}$ for F, where A_{ij} represents the attention weight of the i-th position with respect to the j-th position. Finally, the attention weights are applied to the input X to obtain the multi-scale attention output $O \in R^{n \times d}$, where $O_i = \sum_{j=1}^{n} A_{ij} X_j$. In this way, we obtain a text representation that can capture information at multiple scales, which can be used for subsequent intent classification tasks.

4.7. Loss Function

In this paper, the cross-entropy loss function is adopted as the objective function, as it is commonly used in classification tasks to measure the model's performance on samples and the difference between the probability values of the predicted results and those of the true labels. The following formula represents the cross-entropy loss function:

$$L = -\frac{1}{N} \sum_{i} \sum_{c=1}^{M} y_{ic} \log(p_{ic}), \tag{6}$$

where N denotes the number of samples; M denotes the number of classes; y_{ic} is the indicator function (0 or 1), which takes the value 1 if the actual class of sample i is c; in other cases, it takes the value 0; p_{ic} is the predicted probability of observing sample i belonging to class c. For each sample, calculate the product of the logarithm of the predicted class probability and the true class probability, then take the average over all samples and apply a negative sign. A lower value of the cross-entropy loss function implies that the model's predictions are more aligned with the true labels, indicating better model performance. It can effectively handle the class imbalance problem, giving higher weights to categories with fewer samples. Moreover, the cross-entropy loss function can be optimized using gradient descent, as its derivative with respect to the predicted probability is relatively simple, given by:

$$\frac{\partial L}{\partial p_{ic}} = -\frac{y_{ic}}{p_{ic}}, \tag{7}$$

5. Results

5.1. Dataset

In this work, we evaluate our method using the IMCS-21 [61] dataset, a benchmark corpus for automated medical consultation systems. It gathers authentic online doctor–patient dialogs and provides multi-level manual annotations, including named entities, dialog intents, symptom labels, medical reports, and more. The IMCS-21 dataset contains 4116 samples of doctor–patient conversations with fine-grained annotations. The

specific categories are divided into 16 types, as shown in Table 1, covering 10 pediatric diseases such as pediatric bronchitis, pediatric fever, and pediatric diarrhea. The purpose of the IMCS-21 dataset is to facilitate the advancement of intelligent medical consultation systems and utilize human–computer dialog to assist the consultation process. In the experiment, we carry out the following operations on the data. We split the dataset samples into training, validation, and testing sets in a 6:2:2 ratio, resulting in 2472, 811, and 833 samples, respectively. The training set is used to train the model, the validation set guides optimization decisions, and the test set is employed to evaluate the model's performance and obtain the reported results. Subsequently, the text data undergoes preprocessing, including tokenization, cleaning, and encoding, to generate formats suitable for various model inputs. For Chinese text, we use the Jieba library for tokenization, while for English text, we utilize the NLTK library to ensure consistency in multilingual data processing. The text-cleaning process involves removing special characters, extra spaces, and HTML tags, as well as filtering out stop words. During the input encoding stage, for Transformer-based models such as BERT, we use HuggingFace Tokenizers to encode the text into token IDs and generate attention masks. For traditional models, we employ a vocabulary-based indexing method to ensure compatibility with different architectures. Regarding the class imbalance in the dataset, no additional balancing operations are performed; we directly use the standard cross-entropy loss function.

Table 1. Intent categories and examples in the IMCS-21 medical dialog dataset.

Intent Category	Abbreviation	Example
Request Symptom	R_SX	Patient: Why do I keep coughing?
Inform Symptom	I_SX	Patient: I have a slight fever and sore throat.
Request Etiology	R_ETIOL	Patient: Is the fever due to catching a cold?
Inform Etiology	I_ETIOL	Doctor: It is likely caused by a viral infection.
Request Basic Information	R_BI	Doctor: What is your age?
Inform Basic Information	I_BI	Patient: I am 19 years old.
Request Existing Exam and Treatment	R_EET	Doctor: Did you handle the wound before coming to the hospital?
Inform Existing Exam and Treatment	I_EET	Patient: I did some basic bandaging.
Request Drug Recommendation	R_DR	Patient: What medicine should I take?
Inform Drug Recommendation	I_DR	Doctor: Just take some Lianhua Qingwen capsules.
Request Medical Advice	R_MA	Patient: Do I need to go to the hospital for tests?
Inform Medical Advice	I_MA	Doctor: If the fever lasts more than three days, get a blood test.
Request Precautions	R_PRCTN	Patient: What else should I be aware of?
Inform Precautions	I_PRCTN	Doctor: Rest adequately and avoid catching a chill.
Diagnose	DIAG	Doctor: You have gastritis.
Other	OTR	Patient: Thank you, doctor.

Class Distribution

To improve transparency and support the interpretation of model performance, we analyzed the distribution of dialog acts in the IMCS-21 dataset. Table 2 presents the number of samples per class across the training, validation, and test sets. Figure 4 illustrates the class distribution of the IMCS-21 database. The distribution reveals a certain degree of class imbalance, which poses challenges for classification and highlights the need for robust model design. Note that the "Other" class is excluded from the distribution figure to better highlight the medically meaningful intent categories. Other refers to general or ambiguous utterances that are not directly related to clinical intent in the dataset. In the experiments, the model is still trained and evaluated on all classes, including "Other".

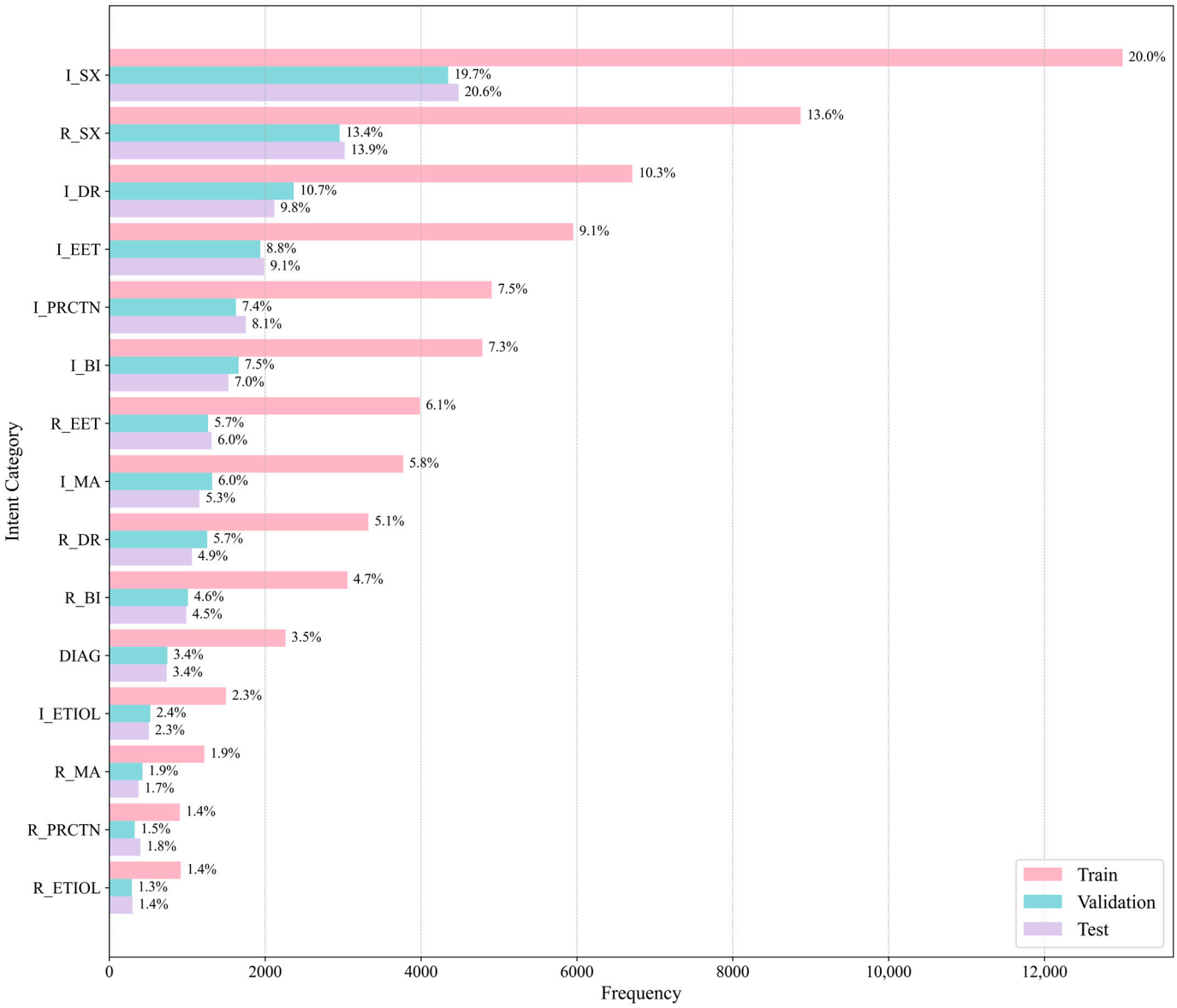

Figure 4. Class distribution diagram of the IMCS-21 database.

Table 2. Number of sample items for the IMCS-21 medical dialog dataset.

Intent Category	Training Set	Validation Set	Test Set
Request Symptom	2957	8871	3020
Inform Symptom	4348	13,001	4481
Request Etiology	289	917	298
Inform Etiology	527	1498	508
Request Basic Information	1010	3056	988
Inform Basic Information	1659	4787	1528
Request Existing Exam and Treatment	1268	3986	1313
Inform Existing Exam and Treatment	1938	5953	1987
Request Drug Recommendation	1256	3325	1061
Inform Drug Recommendation	2366	6712	2118
Request Medical Advice	425	1219	373
Inform Medical Advice	1320	3770	1157
Request Precautions	326	904	400
Inform Precautions	1624	4906	1752
Diagnose	746	2261	737
Other	11,208	33,363	11,214
Total	33,267	98,529	32,935

5.2. Experimental Setup

In our work, the model construction is implemented using the PyTorch framework (version 1.8.1), and the experiments are performed on a single NVIDIA RTX 3060 GPU (Nvidia, Santa Clara, CA, USA).

5.3. Comparative Experiment

In this experiment, we evaluate seven text classification models, including TEXTCNN, BERT-CNN, ALBERT, BERT, ERNIE, K-BERT, and our proposed MSA K-BERT. These models are applied to a medical text classification task, with accuracy, recall, and F_1-score adopted as evaluation metrics. All models are evaluated on the same dataset: IMCS-21. Table 3 provides a summary of the experimental results.

Table 3. Experiment results of text classification models.

Model	Parameters	Accuracy	Recall	F_1
TEXTCNN	2 M	0.720	0.570	0.603
BERT-CNN	110 M	0.750	0.676	0.674
AlBERT	11 M	0.783	0.695	0.714
BERT	110 M	0.788	0.706	0.717
ERNIE	193.7 M	0.794	0.739	0.743
K-BERT (baseline)	340 M	0.807 ± 0.004	0.727 ± 0.006	0.736 ± 0.005
MSA K-BERT	310 M	0.826 ± 0.003	0.794 ± 0.004	0.810 ± 0.003

From Table 3 and Figure 5, it can be observed that the MSA K-BERT model outperforms all other models in all evaluation indicators. Specifically, in Table 3, we perform five repeated experiments for both the baseline model and MSA K-BERT, calculating the mean and standard deviation of accuracy, recall, and F_1-score. The results demonstrate that the observed average performance improvement of our method over the baseline model is not due to random factors but is statistically significant.

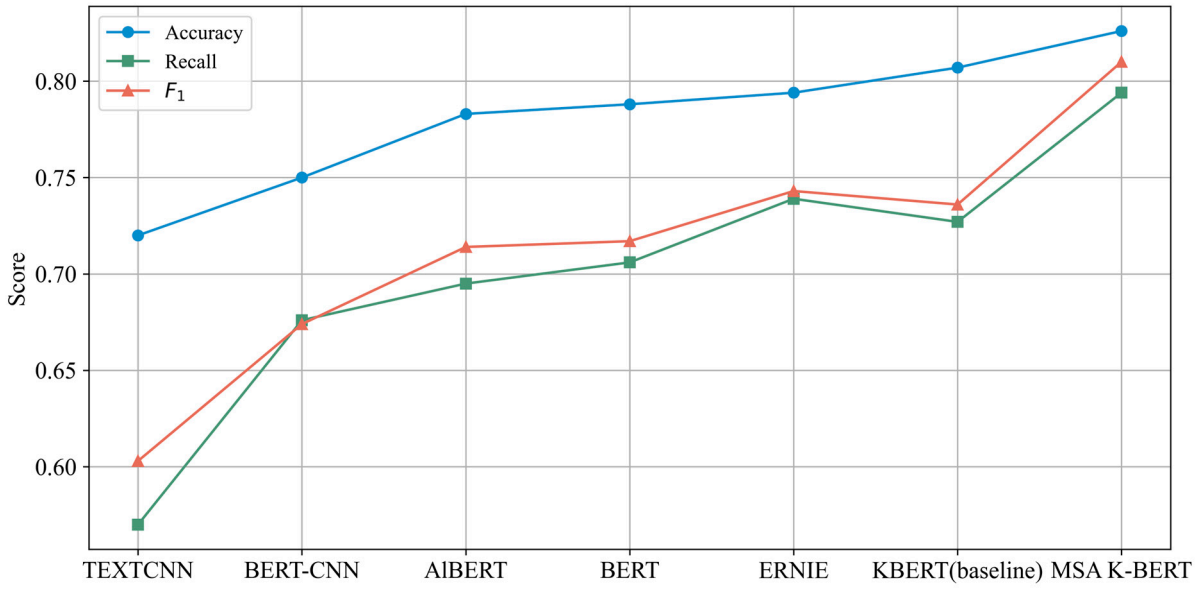

Figure 5. Comparison of the accuracy, recall rate, and F_1 score of the model on the IMCS-21 dataset.

To enhance the statistical robustness of the results, we conducted three additional independent experimental runs of MSA K-BERT. In each run, we split the data into training, validation, and test sets with different ratios while keeping the model architecture consistent. The supplementary experimental results are shown in Table 4. Based on these results, we can conclude that the proposed MSA K-BERT demonstrates significant and robust performance improvements on the IMCS-21 dataset.

Table 4. Experimental data results of MSA K-BERT with different partitioning ratios.

Split Ratio	Accuracy	Recall	F_1
7:2:1	0.819	0.788	0.803
8:1:1	0.814	0.781	0.797
5:3:2	0.808	0.775	0.791

We randomly select 500 samples from the IMCS-21 test set and conduct inference tests for K-BERT and MSA K-BERT under the same hardware environment (NVIDIA RTX 3060 GPU). We measure the average inference time per sample and throughput, with the results shown in Table 5. Although MSA K-BERT may introduce additional attention computations, its actual runtime efficiency still shows certain advantages. In addition, its parameters (310 M) are fewer than those of K-BERT (340 M), indicating that it has the advantages of being higher efficiency and resource-saving. Our model is an improvement based on K-BERT, which is a model that enhances text representation by using KG. This means that our model still has greater potential for improvement. If we can further optimize the selection and use of KG in the model, it may achieve better results.

Table 5. Computational overhead comparison between MSA K-BERT and baseline models.

Model	Parameters	Average Inference Time per Sample (Seconds)	Throughput (Samples/Second)
K-BERT (baseline)	340 M	1.42	0.70
MSA K-BERT	310 M	1.18	0.85

In summary, our model (MSA K-BERT) is an excellent text classification model that outperforms other models in all evaluation metrics, and it also has potential for further

improvement. Our model provides an effective solution for the medical text intent classification task.

5.4. Ablation Study

To verify the validity and superiority of the proposed MSA K-BERT model, we conduct ablation experiments using the K-BERT model as the baseline and test the individual and combined effects of the two main components of our model: multi-scale mechanism (Attention) and MSA. We apply these models to the medical text intent classification task and use accuracy rate, recall rate and F_1 score as evaluation indicators. Table 6 presents the results of the ablation experiment.

From Table 6 and Figure 6, we can see that when we only add the attention component, the accuracy of our model improves, but the recall rate decreases and the F_1 score also increases. This indicates that the attention component helps our model better concentrate on the important information in the text, thereby improving the prediction accuracy. However, it may result in the omission of some relevant information, thereby reducing the coverage of the prediction. When we only add the MSA component, our model shows significant improvements in terms of accuracy, recall rate, and F_1 score, indicating that the MSA component effectively utilizes the multi-head self-attention mechanism to enhance the semantic representation of the text, thereby improving the accuracy and completeness of the prediction. When the attention and MSA components are added simultaneously, our model attains the highest accuracy, recall rate, and F_1 score, indicating that these two components complement each other and synergistically enhance the model's performance.

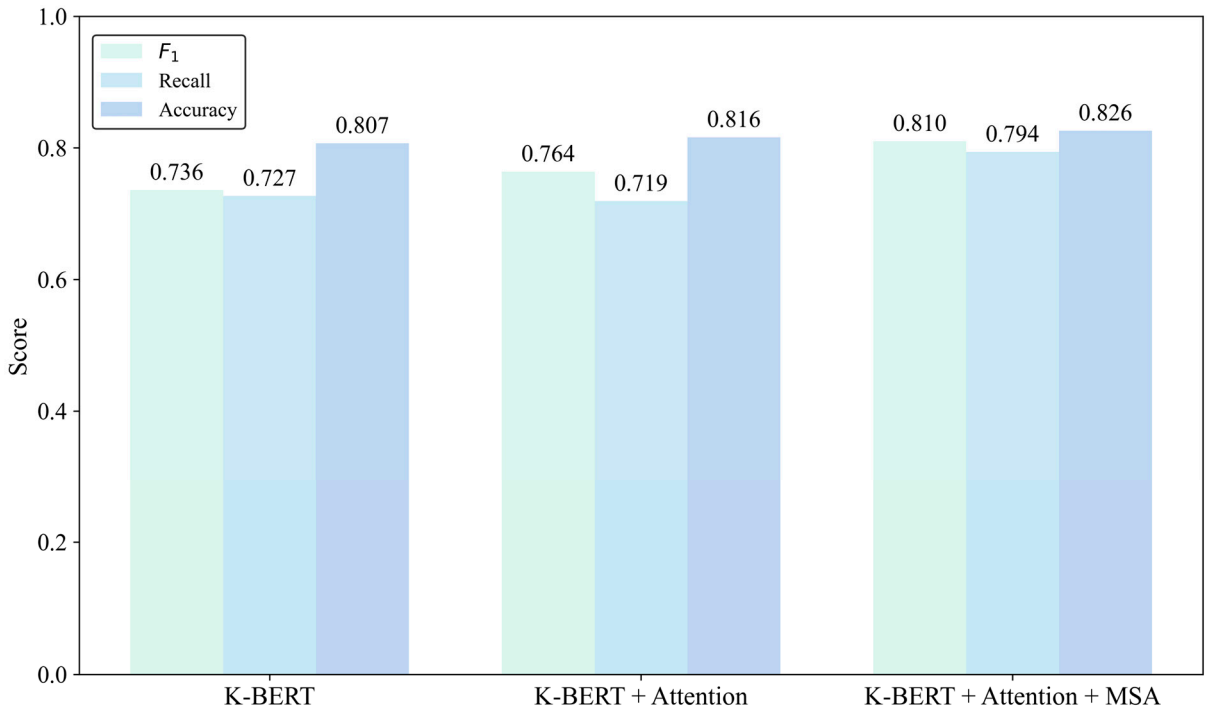

Figure 6. Performance comparison of different configurations.

Through ablation experiments, we systematically evaluate the contribution of each component in the model, demonstrating its overall effectiveness and outstanding performance.

Table 6. The results of the ablation experiments.

Baseline (KBERT)	Attention	MSA	Accuracy	Recall	F_1
★ [1]			0.807	0.727	0.736
★	★		0.816	0.719	0.764
★	★	★	0.826	0.794	0.810

[1] ★ indicates the model components activated in this experimental configuration.

6. Discussion

As one of the core tasks of NLP in smart healthcare [62], medical text intent classification is critical to promoting the accuracy of the diagnostic process, building decision support systems (CDSS) [63], and optimizing doctor–patient interaction. However, existing models still suffer from an insufficient understanding of specialized medical terminology and contextual semantics, as well as ineffective integration of medical KG or domain-specific prior knowledge in practical applications, which weakens semantic support for intent recognition and results in low classification accuracy. Therefore, integrating medical domain knowledge to improve semantic comprehension as well as enhancing the model's generalization ability are problems that need to be resolved in the current research on the intent classification of medical texts.

In this study, we propose a knowledge-enhanced bidirectional encoder representation model that combines multi-scale attention mechanisms. Compared to traditional BERT-based text classification models, MSA K-BERT introduces a BERT-compatible knowledge-supported LR framework that enhances the textual representation with fine-grained medical KG injection strategies, enabling the model to more efficiently capture the implicit or ambiguous semantics commonly found in medical texts. In addition, the introduction of the MSA strengthens the features of different representation layers of our model [60], which significantly promotes the classification accuracy and interpretability of our model. It is worth noting that in our experiments on the IMCS-21 dataset, the accuracy of MSA K-BERT is 0.826, the recall rate is 0.794, and the F_1 score is 0.810. The method we propose achieves superior performance and significantly outperforms both conventional and deep learning-based methods. These results verify the effectiveness of MSA K-BERT in identifying the intent categories of complex medical texts and demonstrate its potential in smart healthcare applications.

However, the computational overhead brought by knowledge injection and multi-scale concern mechanisms may limit the scalability of the model in resource-limited or real-time environments. Noisy or incomplete knowledge may introduce irrelevant information, potentially reducing the model's performance in specific medical scenarios. Another limitation of this study is the absence of direct experimental comparison with large-scale general-purpose language models such as GPT and LLaMA. These models have demonstrated impressive capabilities on a wide range of NLP tasks. However, due to their considerable computational demands, lack of fine-tuned variants for Chinese medical text, and potential data privacy risks, we focus our work on a lightweight and controllable BERT-based framework that is more suitable for domain-specific applications in healthcare. In future work, we plan to explore the integration and comparison of MSA K-BERT with multilingual or general-purpose LLMs under both zero-shot and fine-tuned settings to further understand their performance trade-offs in specialized domains.

Furthermore, we will explore dynamic knowledge selection strategies to further mitigate KN and investigate reducing computational complexity while maintaining performance. Another important future work we are interested in is to explore the extension of the model to support cross-lingual and multi-modal medical data processing, thus

enhancing its robustness and adaptability in different clinical environments to realize superior performance.

7. Conclusions

In this paper, we propose a knowledge-enhanced bidirectional encoder representation model that combines multi-scale attention mechanisms to enhance prediction, aiming to utilize domain knowledge and multi-scale information to promote the accuracy of medical text intent classification. The novelty of this paper is the introduction of the knowledge layer, where relevant triples from KG are inserted into sentences to generate a sentence tree enhanced with semantic information. Meanwhile, to prevent the interference of KG on the sentences, we introduce a seeing layer and use the visibility matrix to limit the visible region of each token, thereby preserving the original order and semantic relationship of the sentences. In the multi-scale attention layer, we use filters of different sizes to convolve the input content and connect the resulting feature maps, and then generate a multiscale attention output to capture information at different scales by calculating the attention weights to enhance the robustness and accuracy of the model. In addition, to better deal with the imbalance problem among different categories of medical intent texts, we optimize the model parameters using the cross-entropy loss function, which improves the generalization ability and interpretability of the model. Through experiments on the publicly available medical text intent classification dataset IMCS-21, MSA K-BERT achieves better performance compared with other advanced deep learning models.

In conclusion, this study provides an effective new method for the medical text intent classification task and offers new perspectives for other similar NLP tasks. However, our model still has potential for enhancement in terms of dataset diversity and computational cost. One limitation of the current study is that all experiments are conducted on Chinese medical texts from the IMCS-21 dataset. While our model demonstrates strong performance in this context, its generalizability to other languages and clinical settings has not yet been evaluated. In the future, we can further expand the scale of the experiment and explore different application scenarios to verify model practicability. We can also reduce the model parameters by using techniques such as lightweight network structures and knowledge distillation to save computational costs.

Author Contributions: Conceptualization, Y.Y.; Methodology, Y.Y.; Writing—original draft preparation, Y.Y.; Writing—review and editing, G.X. and Y.Y.; Supervision, G.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here https://github.com/lemuria-wchen/imcs21(accessed on 20 March 2025). The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We sincerely appreciate the editors and anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, N.; Su, X.; Liu, T.; Hao, Q.; Wei, M. A Benchmark Dataset and Case Study for Chinese Medical Question Intent Classification. *BMC Med. Inf. Decis. Mak.* **2020**, *20*, 125. [CrossRef] [PubMed]
2. Yahia, H.S.; Abdulazeez, A.M. Medical Text Classification Based on Convolutional Neural Network: A Review. *Int. J. Sci. Bus.* **2021**, *5*, 27–41.
3. Yao, L.; Mao, C.; Luo, Y. Clinical Text Classification with Rule-Based Features and Knowledge-Guided Convolutional Neural Networks. *BMC Med. Inf. Decis. Mak.* **2019**, *19*, 71. [CrossRef]
4. Mollaei, N.; Cepeda, C.; Rodrigues, J.; Gamboa, H. Biomedical Text Mining: Applicability of Machine Learning-Based Natural Language Processing in Medical Database. In Proceedings of the Biosignals, Caparica, Portugal, 5 March 2022; pp. 159–166.
5. Lenivtceva, I.; Slasten, E.; Kashina, M.; Kopanitsa, G. Applicability of Machine Learning Methods to Multi-Label Medical Text Classification. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2020; pp. 509–522, ISBN 978-3-030-50422-9.
6. Hassan, S.U.; Ahamed, J.; Ahmad, K. Analytics of Machine Learning-Based Algorithms for Text Classification. *Sustain. Oper. Comput.* **2022**, *3*, 238–248. [CrossRef]
7. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning—Based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2022**, *54*, 1–40. [CrossRef]
8. Shen, Z.; Zhang, S. A Novel Deep-Learning-Based Model for Medical Text Classification. In Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition, Xiamen, China, 30 October 2020; pp. 267–273.
9. Shiri, F.M.; Perumal, T.; Mustapha, N.; Mohamed, R. A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU. *JAI* **2024**, *6*, 301–360. [CrossRef]
10. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 1 June 2019; Volume 1, pp. 4171–4186.
11. Luan, Y.; Lin, S. Research on Text Classification Based on CNN and LSTM. In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 29–31 March 2019; pp. 352–355.
12. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
13. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE Trans. Image Process* **2016**, *25*, 2529–2541. [CrossRef] [PubMed]
14. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. *arXiv* **2014**, arXiv:1404.2188.
15. Zhang, J.; Li, Y.; Tian, J.; Li, T. LSTM-CNN Hybrid Model for Text Classification. In Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 October 2018; pp. 1675–1680.
16. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention Mechanisms in Computer Vision: A Survey. *Comp. Visual. Med.* **2022**, *8*, 331–368. [CrossRef]
17. Zhou, X.; Li, Y.; Liang, W. CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *18*, 912–921. [CrossRef]
18. Liu, Z.; Huang, H.; Lu, C.; Lyu, S. Multichannel CNN with Attention for Text Classification. *arXiv* **2020**, arXiv:2006.16174.
19. Dernoncourt, F.; Lee, J.Y. PubMed 200k RCT: A Dataset for Sequential Sentence Classification in Medical Abstracts. *arXiv* **2017**, arXiv:1710.06071.
20. Tsatsaronis, G.; Balikas, G.; Malakasiotis, P.; Partalas, I.; Zschunke, M.; Alvers, M.R.; Weissenborn, D.; Krithara, A.; Petridis, S.; Polychronopoulos, D.; et al. An Overview of the BIOASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinform.* **2015**, *16*, 138. [CrossRef] [PubMed]
21. Qasim, R.; Bangyal, W.H.; Alqarni, M.A.; Ali Almazroi, A. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *J. Healthc. Eng.* **2022**, *2022*, 1–17. [CrossRef] [PubMed]
22. Zhu, R.; Tu, X.; Huang, J.X. Utilizing BERT for Biomedical and Clinical Text Mining. In *Data Analytics in Biomedical Engineering and Healthcare*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 73–103.
23. Gardazi, N.M.; Daud, A.; Malik, M.K.; Bukhari, A.; Alsahfi, T.; Alshemaimri, B. BERT Applications in Natural Language Processing: A Review. *Artif. Intell. Rev.* **2025**, *58*, 166. [CrossRef]
24. Talaat, A.S. Sentiment Analysis Classification System Using Hybrid BERT Models. *J. Big Data* **2023**, *10*, 110. [CrossRef]
25. Kaur, K.; Kaur, P. BERT-CNN: Improving BERT for Requirements Classification Using CNN. *Procedia Comput. Sci.* **2023**, *218*, 2604–2611. [CrossRef]

26. Li, W.; Gao, S.; Zhou, H.; Huang, Z.; Zhang, K.; Li, W. The Automatic Text Classification Method Based on Bert and Feature Union. In Proceedings of the 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), Tianjin, China, 4–6 December 2019; pp. 774–777.
27. Wang, X.; Bo, D.; Shi, C.; Fan, S.; Ye, Y.; Yu, P.S. A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources. *IEEE Trans. Big Data* **2022**, *9*, 415–436. [CrossRef]
28. Naseem, U.; Thapa, S.; Zhang, Q.; Hu, L.; Masood, A.; Nasim, M. Reducing Knowledge Noise for Improved Semantic Analysis in Biomedical Natural Language Processing Applications. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 272–277.
29. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large Language Models in Medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
30. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774.
31. Mullick, A.; Gupta, M.; Goyal, P. Intent Detection and Entity Extraction from BioMedical Literature. *arXiv* **2024**, arXiv:2404.03598.
32. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
33. Cheng, Y.; Zhao, H.; Zhou, X.; Zhao, J.; Cao, Y.; Yang, C.; Cai, X. A Large Language Model for Advanced Power Dispatch. *Sci. Rep.* **2025**, *15*, 8925. [CrossRef]
34. Wu, J.; Zhou, S.; Zuo, S.; Chen, Y.; Sun, W.; Luo, J.; Duan, J.; Wang, H.; Wang, D. U-Net Combined with Multi-Scale Attention Mechanism for Liver Segmentation in CT Images. *BMC Med. Inf. Decis. Mak.* **2021**, *21*, 283. [CrossRef] [PubMed]
35. Hogan, A.; Blomqvist, E.; Cochez, M.; D'amato, C.; Melo, G.D.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S.; et al. Knowledge Graphs. *ACM Comput. Surv.* **2022**, *54*, 1–37. [CrossRef]
36. Sengupta, S.; Krone, J.; Mansour, S. On the Robustness of Intent Classification and Slot Labeling in Goal-Oriented Dialog Systems to Real-World Noise. *arXiv* **2021**, arXiv:2104.07149.
37. Kalinowski, A.; An, Y. A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs. *arXiv* **2020**, arXiv:2010.13688.
38. Khalid, B.; Dai, S.; Taghavi, T.; Lee, S. Label Supervised Contrastive Learning for Imbalanced Text Classification in Euclidean and Hyperbolic Embedding Spaces. In Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024), San Ġiljan, Malta, 15 November 2024; pp. 58–67.
39. Mitra, V.; Wang, C.-J.; Banerjee, S. Text Classification: A Least Square Support Vector Machine Approach. *Appl. Soft Comput.* **2007**, *7*, 908–914. [CrossRef]
40. Isa, D.; Lee, L.H.; Kallimani, V.P.; Rajkumar, R. Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1264–1272. [CrossRef]
41. Wu, Q.; Ye, Y.; Zhang, H.; Ng, M.K.; Ho, S.-S. ForesTexter: An Efficient Random Forest Algorithm for Imbalanced Text Categorization. *Knowl.-Based Syst.* **2014**, *67*, 105–116. [CrossRef]
42. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123*, 8736–8780. [CrossRef] [PubMed]
43. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *Sn Comput. Sci.* **2021**, *2*, 420. [CrossRef] [PubMed]
44. Liu, G.; Guo, J. Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification. *Neurocomputing* **2019**, *337*, 325–338. [CrossRef]
45. Wang, J.; Yu, L.-C.; Lai, K.R.; Zhang, X. Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis. *IEEE/ACM Trans. Audio Speech Lang. Process* **2019**, *28*, 581–591. [CrossRef]
46. Han, Y.; Guo, J.; Yang, H.; Guan, R.; Zhang, T. SSMA-YOLO: A Lightweight YOLO Model with Enhanced Feature Extraction and Fusion Capabilities for Drone-Aerial Ship Image Detection. *Drones* **2024**, *8*, 145. [CrossRef]
47. Liu, X.; Wang, Z.; Han, Y.; Wang, Y.; Yuan, J.; Song, J.; Zheng, B.; Zhang, L.; Huang, S.; Chen, H. Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models. *arXiv* **2025**, arXiv:2501.05179.
48. Han, Y.; Duan, B.; Guan, R.; Yang, G.; Zhen, Z. LUFFD-YOLO: A Lightweight Model for UAV Remote Sensing Forest Fire Detection Based on Attention Mechanism and Multi-Level Feature Fusion. *Remote Sens.* **2024**, *16*, 2177. [CrossRef]
49. Liu, X.; Liu, T.; Huang, S.; Xin, Y.; Hu, Y.; Qin, L.; Wang, D.; Wu, Y.; Chen, H. M2IST: Multi-Modal Interactive Side-Tuning for Efficient Referring Expression Comprehension. IEEE Transactions on Circuits and Systems for Video Technology. *arXiv* **2025**, arXiv:2407.01131v3.
50. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv* **2019**, arXiv:1904.09223.
51. Clark, K.; Luong, M.-T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators. *arXiv* **2020**, arXiv:2003.10555.

52. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv* **2020**, arXiv:1909.11942.
53. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv* **2016**, arXiv:1606.05250.
54. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv* **2019**, arXiv:1804.07461.
55. Moradi, M.; Blagec, K.; Samwald, M. Deep Learning Models Are Not Robust against Noise in Clinical Text. *arXiv* **2021**, arXiv:2108.12242.
56. Si, Y.; Wang, J.; Xu, H.; Roberts, K. Enhancing Clinical Concept Extraction with Contextual Embeddings. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1297–1304. [CrossRef]
57. Zhang, H.; Lu, A.X.; Abdalla, M.; McDermott, M.; Ghassemi, M. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. In Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto, ON, Canada, 2 April 2020; pp. 110–120.
58. Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; Wang, P. K-Bert: Enabling Language Representation with Knowledge Graph. *AAAI Conf. Artif. Intell.* **2020**, *34*, 2901–2908. [CrossRef]
59. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
60. Chen, W.; Shi, K. Multi-Scale Attention Convolutional Neural Network for Time Series Classification. *Neural Netw.* **2021**, *136*, 126–140. [CrossRef]
61. Chen, W.; Li, Z.; Fang, H.; Yao, Q.; Zhong, C.; Hao, J.; Zhang, Q.; Huang, X.; Peng, J.; Wei, Z. A Benchmark for Automatic Medical Consultation System: Frameworks, Tasks and Datasets. *Bioinformatics* **2023**, *39*, btac817. [CrossRef]
62. Tian, S.; Yang, W.; Le Grange, J.M.; Wang, P.; Huang, W.; Ye, Z. Smart Healthcare: Making Medical Care More Intelligent. *Glob. Health J.* **2019**, *3*, 62–65. [CrossRef]
63. Bright, T.J.; Wong, A.; Dhurjati, R.; Bristow, E.; Bastian, L.; Coeytaux, R.R.; Samsa, G.; Hasselblad, V.; Williams, J.W.; Musty, M.D.; et al. Effect of Clinical Decision-Support Systems: A Systematic Review. *Ann. Intern Med.* **2012**, *157*, 29–43. [CrossRef] [PubMed]