Image-Based Contextual Pill Recognition with Medical Knowledge Graph Assistance

Anh Duy Nguyen[1], Thuy Dung Nguyen[1], Huy Hieu Pham[2,3], Thanh Hung Nguyen[1], and Phi Le Nguyen[1(✉)]

[1] Hanoi University of Science and Technology, Hanoi, Vietnam
lenp@soict.hust.edu.vn
[2] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[3] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam

Abstract. In many healthcare applications, identifying pills given their captured images under various conditions and backgrounds has been becoming more and more essential. Several efforts have been devoted to utilizing the deep learning-based approach to tackle the pill recognition problem in the literature. However, due to the high similarity between pills' appearance, misrecognition often occurs, leaving pill recognition a challenge. To this end, in this paper, we introduce a novel approach named PIKA that leverages external knowledge to enhance pill recognition accuracy. Specifically, we address a practical scenario (which we call contextual pill recognition), aiming to identify pills in a picture of a patient's pill intake. Firstly, we propose a novel method for modeling the implicit association between pills in the presence of an external data source, in this case, prescriptions. Secondly, we present a walk-based graph embedding model that transforms from the graph space to vector space and extracts condensed relational features of the pills. Thirdly, a final framework is provided that leverages both image-based visual and graph-based relational features to accomplish the pill identification task. Within this framework, the visual representation of each pill is mapped to the graph embedding space, which is then used to execute *attention* over the graph representation, resulting in a semantically-rich context vector that aids in the final classification. To our knowledge, this is the first study to use external prescription data to establish associations between medicines and to classify them using this aiding information. The architecture of PIKA is lightweight and has the flexibility to incorporate into any recognition backbones. The experimental results show that by leveraging the external knowledge graph, PIKA can improve the recognition accuracy from 4.8% to 34.1% in terms of *F1*-score, compared to baselines.

Keywords: Pill recognition · Knowledge graph · Graph embedding

1 Introduction

Medicines are used to cure diseases and improve patients' health. Medication mistakes, however, may have serious consequences, including diminishing the efficacy of the treatment, causing adverse effects, or even leading to death. According

E. Szczerbicki et al. (Eds.): ACIIDS 2022, CCIS 1716, pp. 354–369, 2022.
https://doi.org/10.1007/978-981-19-8234-7_28

Fig. 1. Ill-predicted medicines

to a WHO report, one-third of all mortality is caused by the misuse of drugs, not by disease [2]. Moreover, according to Yaniv *et al.* [20], medication errors claim the lives of about six to eight thousand people every year. To emphasize the significance of taking medication correctly, WHO has chosen the subject Medication Without Harm for World Patient Safety Day 2022 [1].

Medication errors may fall into many categories, one of which is incorrect pill intake, which occurs when the drugs taken differ from the prescription. This is due to the difficulty in manually distinguishing pills owing to the wide variety of drugs and similarities in pill colors and shapes. In such a context, many attempts have been made to assist users in identifying the pills automatically. In recent years, machine learning (ML) has emerged as a viable technique for tackling object classification problems. Many studies have employed machine learning in the pill recognition challenge [3,15,19]. Some common techniques such as convolutional neural networks (CNN) and Graph Neural Networks (GNN) are often used. For instance, in [19], the authors exploited Deep Convolution Network to identify pills. In [15], Enhanced Feature Pyramid Networks (EFPNs) and Global Convolution Network (GCN) are combined to enhance the pill localization accuracy. Besides, the authors leveraged the Xception network [4] to solve the pill recognition problem. The authors in [3] studied how to help visually impaired chronic patients in taking their medications correctly. To this end, they proposed a so-called MedGlasses system, which combines AI and IoT. MedGlasses comprises smart glasses capable of recognizing pills, a smartphone app capable of reading medication information from a QR code and reminding users to take the medication, and a server system to store user information. Furthermore, numerous efforts have strived to improve pill recognition accuracy by incorporating handcrafted features such as color, shape, and imprint. Ling et al. [9] investigated the problem of few-shot pill detection. The authors proposed a Multi-Stream (MS) deep learning model that combines information from four streams: RGB, Texture, Contour, and Imprinted Text. In addition, they offered a two-stage training technique to solve the data scarcity constraint; the first stage is to train with all samples, while the second concentrates only on the hard examples. In [12], the authors integrated three handcrafted features, namely shape, color, and imprinted text, to identify pills. Specifically, the authors first used statistical measurements from the pill's histogram to estimate the number of colors in the pill. The imprinted text on the pill was then extracted using text recognition tools. The author also used the decision tree technique to determine the pill shape. The color, shape, and imprinted text information are then used as input features to train the classification model.

Despite numerous efforts, pill recognition remains problematic. Especially, pill misidentification often occurs with tablets that look substantially similar. Figure 1 shows some of the misclassification results made by a deep learning model. To overcome the limitations of existing approaches, in this study, we propose a novel method that leverages external knowledge to increase the accuracy and, in particular, to tackle the misclassification of similar pills. Unlike the existing works, we focus on a practical application that recognizes pills in the patient's pill intake picture. The external knowledge we use is the information extracted from a given set of prescriptions. Our main idea is that by using such external knowledge, we can learn the relationship between the drugs, such as the co-occurrence likelihood of the pills. This knowledge will be utilized to improve the pill recognition model's accuracy.

To summarize, our main contributions are as follows:

- We are the first to address a so-called *contextual pill recognition* problem, which recognizes pills in a picture of a patient's pill intake.
- We build a dataset containing pill images taken in unconstraint conditions and a corresponding prescription collection.
- We propose a novel deep learning-based approach to solve the contextual pill recognition problem. Specifically, we design a method to construct a prescription-based knowledge graph representing the relationship between pills. We then present a graph embedding network to extract pills' relational features. Finally, we design a framework to fuse the graph-based relational information with the image-based visual features to make the final classification decision.
- We design loss functions and a training strategy to enhance the classification accuracy.
- We conduct thorough experiments on a dataset of drugs taken in real-world settings and compare the performance of the proposed solution to existing methods. The experimental findings indicate that our proposed model outperforms significantly the baselines.

The remainder of the paper is organized as follows. We introduce the related works in Sect. 2. Section 3 describes our proposed solution. We evaluate the effectiveness of the proposed approach in Sect. 4 and conclude the paper in Sect. 5. Our code and pre-trained deep learning models will be made publicly available on our project webpage (http://vaipe.io/) upon the publication of this paper.

2 Related Work

The contextual pill recognition can be treated as a traditional object identification problem. The conventional approach is to divide it into two stages. The first stage is responsible for segmenting each pill image, and the second one treats each pill box as a separate object and identifies it using an object recognition

model. In [19], the authors employed Deep Convolution Network, Feature Pyramid Networks (EFPNs), combined with GCN for pill detection. They then used the Xception network to identify the pill. Ling et al. [9] studied the issue of pill identification with a limited number of samples. To improve identification accuracy, the author incorporated data from numerous sources, including RGB, Texture, Contour, and Imprinted Text. Hand-crafted features such as shape, color, and imprinted text were also used in the [12]. The shortcoming of these approaches is that they handle each pill in the picture separately, without taking use of the pill's interaction.

Contextual pill identification is analogous to the multi-label classification problem, which has drawn a lot of attention in recent years. Many research has employed external information to improve recognition accuracy in this problem. The most common strategy is to obtain the label co-occurrence relationship and use it in the recognition task. Label co-occurrence may be retrieved using a variety of approaches, including probabilistic models, neural networks, and graph networks. Li et al. employed conditional graphical Lasso model in [8] to statistically calculate the co-occurrence probability of the labels. Several works have adopted neural networks such as LSTM to simulate the interaction of labels to decrease the computation costs [16]. The author of [5] used the autoencoder Graph Isomorphism Network (GIN) to represent the label association. They also presented a collaborative training framework incorporating label semantic encoding and label-specific feature extraction. There are also several techniques to utilize relational information. Relational information, in particular, may be combined with visual characteristics in the final layers, as shown in [18]. It may also be injected into the middle CNN layers through lateral connections, as described in [17].

Contextual pill identification, on the other hand, differs from traditional multi-label classification in that the multi-label classification task tries to recognize the global information offered by the picture rather than finding and recognizing each item featured in the image. The second problem lies in modeling the label's relation. Indeed, traditional multi-label classification systems mainly construct label relationships based on the semantic meaning of the label's name. This strategy, however, does not work with medicine names since they often have no meaning. Furthermore, extracting correlations between medicine names from public data sources is difficult.

3 Proposed Approach

In this section we propose a novel pill recognition framework named PIKA (which stands for **P**ill **I**dentification with medical **K**nowledge gr**A**ph). We first present the main components of the PIKA framework (Sect. 3.1). We then describe how the prescription-based medical knowledge graph is built (Sect. 3.2) and explain how pill visual features are extracted (Sect. 3.3). Next, we combine the built medical knowledge graph and extracted visual features to enhance pill identification performance (Sect. 3.4). Finally, we introduce an auxiliary loss and a training strategy to improve the effectiveness of the proposed learning model (Sect. 3.5).

3.1 Overview

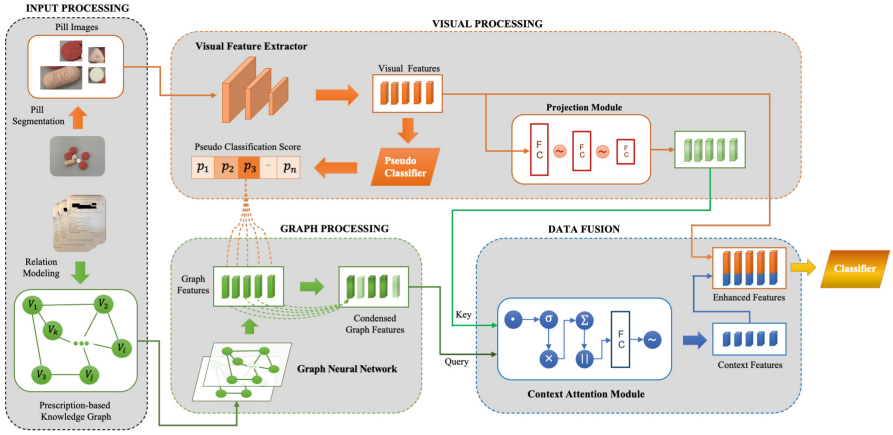

Fig. 2. Overview of the Proposed Framework. Firstly, the Input Processing Procedure is used to generate a non-directed Medical Knowledge Graph (MKG) $\mathcal{G} = <\mathcal{V}, \mathcal{E}, \mathcal{W}>$ from given prescriptions, and crop the input images into pill boxes. Secondly, the MKG is fed into a Graph embedding network to extract pills' relational features, while the cropped pill images are passed via a backbone network to retrieve its visual representations. At this stage, the graph-based relational features are combined with the pseudo-class scores produced by the visual backbone to make up its condensed version. Thirdly, the visual embedding get projected to the same hyper-plane as their counterparts in the graph space, with the aid of the Projection Module. Following, the projected visual features, coupled with graph-based relational information, are the input for the Context Attention Module to provide the context vector. Finally, the enriched visual features, which combine the context vector and the visual features, are fed into the final classifier to identify the pill.

As illustrated in Fig. 2, the proposed model comprises four major components: input processing, visual processing, graph processing, and information fusion. The first block, i.e., input processing, is in charge of locating and retrieving pill images and creating a graph modeling drug interactions. The visual processing block is used to extract visual features of the pills, while the graph processing module attempts to depict the relationship between the pills. The fusion layer then combines the visual characteristics of the pills with their graph-based relational features to generate the final classification decision. The overall flow is as follows.

- **Step 1**. The original image containing multiple pills is passed through an object localization model to identify and cut out bounding-box images of every pill. Note that we do not focus on the object detection problem in this work; thus, we can use any object detection model for this step.
- **Step 2**. We construct a graph from a given set of prescriptions, with nodes representing pills and edges reflecting drug linkages. We name this graph

the Prescription-based Medical Knowledge Graph or PMKG for short. The PMKG is then passed through a Graph Neural Network (GNN) to yield embedding vectors. Each embedding vector conveys information about a node and its relationship to the neighbors. The detailed algorithm is presented in Sect. 3.2.

- **Step 3**. The pills' images will then be put into the Visual processing module to extract the visual characteristic. On the one hand, these features will be fed into the data fusion block to make the classification decision. On the other hand, these features are put in a projection module. The objective of the projection module is to generate a representation similar to that of the graph processing block. The projected features are then utilized to learn the relationship between the pill images and the PMKG's nodes. The detail of the visual processing module will be described in Sect. 3.3.
- **Step 4**. The Graph embedding vector retrieved in Step 2 and the projected features acquired in Step 3 will be passed through an attention layer to generate a context vector. Finally, the context vector will be concatenated with the Visual feature before being fed into the final classifier, which will produce prediction results. The details of our losses functions are presented in Sect. 3.5.

3.2 Prescription-Based Medical Knowledge Graph

The key idea behind the proposed approach is to utilize the information on the relationship between pills via their corresponding prescriptions to enhance image-based pill recognition. To this end, a prescription-based medical knowledge graph is constructed. Our intuition is that all the medicines are prescribed to cure or alleviate some diseases or symptoms in actual pill captures. Hence, we can formulate that implicit relation through the direct relations between pills and diagnoses. This information contains in the prescriptions provided by pharmacists to their patients. This section covers our detailed methodology for knowledge graph modeling and our framework for embedding this graph.

Knowledge Graph Modelling. The Medical Knowledge Graph (MKG) is a weighted graph, denoted as $\mathcal{G} = <\mathcal{V}, \mathcal{E}, \mathcal{W}>$, whose vertices $\mathcal{V}$ represent pill classes, and the weights $\mathcal{W}$ indicate the relationship between the pills. With prescriptions as the initial data, two factors can be used to formulate graph edges $\mathcal{E}$, which are diagnoses and medications. As the relationship between pills is not explicitly presented in prescriptions, we model the relation representing the edge between two nodes (i.e., pill classes) C_i and C_j based on the following criteria.

- There is an edge between two pill classes C_i and C_j if and only if they have been prescribed for at least one shared diagnosis.
- The weight of an edge E_{ij} connecting pill classes C_i and C_j reflects the likelihood that these two medications will be given at the same time.

Instead of directly weighting the `Pill-Pill` edges, we determine the weights via `Diagnose-Pill` relation. In particular, we first define a so-called `Diagnose-Pill` impact factor, which reflects how important a pill is to a diagnosis or, in other words, how often a pill is prescribed to cure a diagnosis. Inspired by the Term Frequency (`tf`) – Inverse Dense Frequency (`idf`) often used in NLP domain, we define the impact factor of a pill P_j to a diagnose D_i (denoted as $I(P_j, D_i)$) as follows.

$$\mathcal{I}(P_j, D_i) = \texttt{tf}(D_j, P_i) \times \texttt{idf}(P_i) = \frac{|\mathbb{S}(D_j, P_i)|}{|\mathbb{S}(D_j)|} \times \log \frac{|\mathbb{S}|}{|\mathbb{S}(P_i)|}, \tag{1}$$

where $\mathbb{S}$ represents the set of all prescriptions, $\mathbb{S}(D_j, P_i)$ depicts the collection of prescriptions containing both D_j and P_i, and $\mathbb{S}(D_j)$ illustrates the set of prescriptions containing D_j. After calculating the impact factors of the pills and diagnoses, we derive the weights between two pills by averaging their impact factors against all diagnoses, as shown below

$$\mathcal{W}(P_i, P_j) = \sum_{D \in \mathbb{D}} \mathcal{I}(P_i, D) + \mathcal{I}(P_j, D), \tag{2}$$

where $\mathcal{W}(P_i, P_j)$ depicts the weight between pills P_i, P_j, and $\mathbb{D}$ denotes the set of all diagnoses.

Knowledge Graph Embedding. As the MKG is sparse, we will not utilize it directly but pass it through a graph embedding module to extract the condensed meaningful information. Specifically, the graph embedding module helps project from the graph space into a vector space. Each vector corresponds to a node (i.e., a pill class) and conveys information about that node and its relationships with the neighbors. With the graph embedding module, we want to preserve the co-occurrence property of the pills, i.e., if two nodes V_i, V_j are neighbors in the original MKG, their corresponding presentations u_i and u_j should also have small distance in the vector space. To this end, we leverage the walk-based approach, which will train a graph embedding network using the skip-gram model with the following loss function

$$\mathcal{L}_g = -\sum_{i=1}^{n} \left[\sum_{u_j \in \mathbb{N}(u_i)} \sigma\left(u_i \cdot u_j\right) - \sum_{u_k \notin \mathbb{N}(u_i)} \log\left(e^{u_k \cdot u_i}\right) \right], \tag{3}$$

where n denotes the total number of nodes in the graph, u_i represents the embedding vector of node V_i, and $\mathbb{N}(u_i)$ depicts the set of V_i's 1-hop neighbors. By minimizing $\mathcal{L}_g$, we can reduce the distance between representations of neighboring nodes while increasing that of non-neighboring nodes.

3.3 Visual Processing Procedure

The Visual Processing block is responsible for retrieving two types of information. The former refers to the visual characteristics of individual pill images;

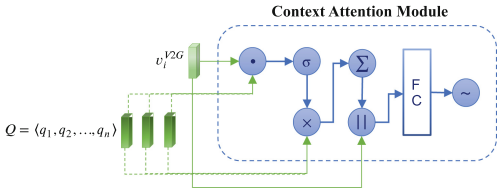

Fig. 3. Illustration of the Context Attention Module

meanwhile, the latter relates to the relational feature that represents the interaction between pills. We employ a conventional Convolutional neural networks such as VGG [14] or ResNet [6] to extract the visual features. Concerning the relational features, our idea is to distill knowledge from MKG into the representation of each pill using a projection layer. Besides, the Visual Processing block also contains a pseudo classifier module that helps to provide rough classification results. These results are then used to filter out condensed information from MKG (the details will be presented in Sect. 3.4).

V2G Projection Layer. The V2G projection layer obtains a pill's visual feature vector as the input. It passes through several layers to generate a representation with the same dimension as the MKG's node embedding vector. This layer can be mathematically represented as $v_i^{V2G} = \theta(v_i)$, where v_i and v_i^{V2G} are the representations in the visual and graph spaces, respectively; and $\theta(\cdot) : V \rightarrow U$ is a non-linear mapping. In the implementation, we formulate this mapping as a stack of Fully Connected (FC) layers, with *tanh* as the middle activation function. Through the training processing, the $\theta(\cdot)$ will be optimized so that the probability distribution of the projected vectors is similar to that of the MKG's embedding vectors. This is accomplished by introducing the Linkage loss as described in Sect. 3.5.

Pseudo Classification Layer. The pseudo classifier produces temporary identification results. This rough classification result will be used as a filter layer responsible for extracting from the MKG only information related to the pills in the image (and omitting information from the nodes that are not associated with the pills in the picture). In our implementation, pseudo classification is currently implemented as a fully connected layer.

3.4 Data Fusion

In the Data Fusion block, we first extract the condensed information from the MKG and integrate it with the visual features using an attention mechanism to create the context feature. The context feature are then concatenated with the visual features before being fed into the final classification layer to make the final decision.

Condensed Relational Feature Extraction. The idea of the Condensed Feature Extraction module is to extract from the MKG only information related to the pills in the input image. Let N be the number of pill classes and M be the number of pills in the input image. Suppose $P = [p_{ij}]_{M \times N}$ is the matrix whose row vectors represent the logits produced by the pseudo classifier, and $U = [u_{kl}]_{N \times H}$ denotes the embedding matrix, whose each row represents a pill class' embedding vector. The condensed relational features, denoted as $\mathcal{R}$ is a set of M vectors, each depicts the condensed relational information of a pill (in the input image), extracted from the MKG. $\mathcal{R}$ is calculated by multiplying the softmax of P to U as follows.

$$\mathcal{R} = \sigma(P) \cdot U. \tag{4}$$

Here the symbol σ denotes the `Softmax` activation function. Intuitively, $\mathcal{R}$ is a matrix consisting of M rows. The i-th row of $\mathcal{R}$ is a weighted sum of all the MKG's node embeddings, whose weights are the classification probabilities corresponding to the i-th pill in the input image.

Attention Module. To avoid misclassification and improve the model's accuracy, besides the pure visual information extracted by the visual extractor described in Sect. 3.3, we leverage the attention mechanism to create a context vector that integrates both visual and relational features. The details of the attention module are illustrated in Fig. 3. Specifically, we use the projected features (i.e., produced by the Visual Projection module) as the key and value, and the graph embedding vectors as the query. The attention module first calculates the attention weights as the similarity of each projected feature to all the graph embedding vectors. The final context vector is then generated by taking the weighted sum of the projected features. The context vector is then fused with the visual features and passed to the final classifier.

3.5 Loss Functions

With additional modules for different purposes mentioned above, we also provide auxiliary losses for aiding the optimization process of the modules. This section is dedicated for presenting those losses.

Classification Loss. Our first objective function is the classification loss, which is composed of two terms. The first term's functionality is to bring the final output of PIKA's Classifier close to the actual result. The following one deal with the output of the Pseudo Classification layer. It is expected to produce the result that is best closed to ground truth also. Since we are dealing with multi-label problems, cross-entropy is used.

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(\tilde{y}_i\right) - \beta_{\text{loose}} * \frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p_i\right), \tag{5}$$

where y_i denotes the one-hot vector - the ground truth of the i- pill, $\tilde{y}_i$ and p_i represent the Classifier's and the Pseudo Classifier's outputs, respectively. In the formula, there is an additional parameter β_{loose} $(0 \leq \beta_{\text{loose}} \leq 1)$, which helps loosing the constraints for the Pseudo Classifier. The closer β_{loose} approaches 1, the more we expect that the output of Pseudo Classifier is similar to the main one. In our case, it is set as 0.1 for additional flexibility.

Linkage Loss. For the V2G projection layer, we propose an auxiliary loss called linkage loss, helping this module achieve its aim, i.e., mapping from the visual space $\mathcal{V}$ to the graph space $\mathcal{U}$. Let $\mathcal{V}^{V2G}$ be the vector spaces produced by the Projection layer; then, our objective is to bring $\mathcal{V}^{V2G}$ close to the graph space $\mathcal{U}$. To this end, we design the linkage loss as a type of probabilistic distance instead of a point-wise one. We believe it would loosen the constraint while also being robust enough to help the module converge. Let the distributions of $\mathcal{V}^{V2G}$ and $\mathcal{U}$ be modeled by two continuous random variables X, and G respectively. Firstly, we have to model the geometry of the distributions. A common way to accomplish this purpose is to investigate the pairwise interactions between sample points (with an ample number of data samples) [7,11] via the joint probability density of every two data samples. Let u_{ij} and v_{ij} be the joint density probability functions of the i-th and j-th data points of variable X and G, respectively; then u_{ij} and v_{ij} can be modeled using Kernel Density Estimation (KDE) [13] as follows.

$$u_{ij} = u_{i|j}u_j = \frac{1}{N}K\left(\mathbf{g}_i, \mathbf{g}_j; 2\sigma_{\mathcal{U}}^2\right); \quad v_{ij}^{V2G} = v_{i|j}^{V2G}v_j^{V2G} = \frac{1}{N}K\left(\mathbf{x}_i, \mathbf{x}_j; 2\sigma_{\mathcal{V}^{V2G}}^2\right), \tag{6}$$

in which $K\left(\cdot, \cdot; 2\sigma^2\right)$ denotes a symmetric kernel function with the width σ; $\mathbf{g}_i, \mathbf{g}_j$ are two data points sampled from the distribution of G, and $\mathbf{x}_i, \mathbf{x}_j$ are two data points sampled from the distribution of X; N is the number of samples. Ideally, we want to minimize the divergence of the joint density probability functions of U and V^{V2G}. However, learning a projection module that can accomplish this purpose is impossible. To alleviate this issue, we choose to replace the joint probability density function with the conditional probability distribution of the samples. Though the divergence of both two functions has the same convergence point (in case the kernel similarities are the same for both distributions), the use of conditional probability can better describe the local regions between data points [11] (expresses the probability of each sample to select each of its neighbors). The conditional probability distributions for the graph and projected visual spaces are defined as follows.

$$u_{i|j} = \frac{K\left(\mathbf{g}_i, \mathbf{g}_j; 2\sigma_{\mathcal{U}}^2\right)}{\sum_{k=1,k\neq j}^{N} K\left(\mathbf{g}_k, \mathbf{g}_j; 2\sigma_{\mathcal{U}}^2\right)}; \quad v_{i|j}^{V2G} = \frac{K\left(\mathbf{x}_i, \mathbf{x}_j; 2\sigma_{V^{V2G}}^2\right)}{\sum_{k=1,k\neq j}^{N} K\left(\mathbf{x}_k, \mathbf{x}_j; 2\sigma_{V^{V2G}}^2\right)}. \tag{7}$$

We use the Cosine Similarity as the kernel. Finally, the divergence metric we chose as our linkage loss function is the Jensen-Shannon (JS) Divergence:

$$\mathcal{L}_l = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}\left[u_{j|i}\log\left(\frac{u_{j|i}}{v_{j|i}^{V2G}}\right) + v_{j|i}^{V2G}\log\left(\frac{v_{j|i}^{V2G}}{u_{j|i}}\right)\right]. \tag{8}$$

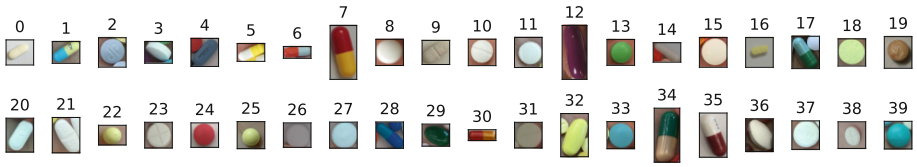

Fig. 4. The visualization of several representative examples from our customized pill dataset.

The total loss comprises of the Classification loss and Linkage loss as follows:

$$\mathcal{L} = \alpha\mathcal{L}_c + (1-\alpha)\mathcal{L}_l, \text{with } \alpha \in (0,1)\,. \tag{9}$$

4 Performance Evaluation

In this section we evaluate the performance of our proposed model, PIKA. We perform several experiments on our custom pill images captured with mobile phones under unconstraint environments. Details about the dataset, together with our evaluation metrics and implementation details, would be covered in Subsect. 4.1. The numerical results are then presented in Sect. 4.2 and 4.3.

4.1 Experimental Setup

Dataset. To the best of our knowledge, the dataset of pill images and corresponding prescription set are not publicly available. That is our motivation to build our own dataset for this work. Table 1 describes some important statistics about it. In addition, the collecting and processing procedure is combined by the following steps.

Table 1. Dataset statistics

Images	Classes	Prescriptions	Train set	Test set
3,087	76	168	116 prescriptions 2,058 images	52 prescriptions 1,029 images

- We collect anonymous prescriptions of 168 patients from 4 hospitals in Vietnam. After processing the raw data, we converted them into JSON format for each prescriptions record; the pills are also indexed to form a dictionary, including 76 kinds of drugs.
- Since the process of collecting pills in accordance with prescriptions takes a great effort, time, and funding, in this current work, we have to collect images of 76 type of drugs which is not exactly the types in our collected prescriptions.

- Following, the collected pills are relabeled by our pill dictionary described above and grouped by prescriptions, with the number of images per prescription being 5. Figure 4 illustrates the appearances of collected pills with their mapped labels.
- Finally, we combine the retrieved sets of prescription photos into two sets, one for the training process and one for evaluation.

Evaluation Metrics. For assessing PIKA performance across all used backbones, as well as other testing scenarios, *Recall*, *Precision* and $F1$-score metrics are adopted altogether. The figures claimed in Sect. 4.2 and Sect. 4.3 are the averaged numbers achieved over all classes.

Implementation Details. In our PIKA implementation, the dimensions of the graph embeddings are set as 64. The Projection Module consists of 3 Fully Connected (FC) Layers, with middle *tanh* activation, and the output dimensions are 512, 256, 64 respectively. The optimizer used is AdamW [10], and the initial learning rate is 0.001. β_{loose} (Eq. 5) and α (Eq. 9) are set as 0.1 and 0.9 respectively. During the training process, the input images is resized to 224×224, with random rotation of 10° and horizontal flip for augmentation. The batch size is set as 32. For the backbones we fuse our framework with, all are kept the same as in the original papers [6,14], except the last classifier is adopted to output 76 scores in compliant with our 76 classes. All the implementation is performed with the help of *Pytorch* framework, and the training, as well as evaluation processes, are conducted with 2 NVIDIA GeForce RTX 3090 GPUs.

We use a two-step training approach for PIKA. We first train the graph module to convergence with its specified loss (Eq. 3). The converged model output is then utilized to train the rest of the PIKA framework. By doing so, we ensure that the graph embeddings are reliable enough and truly reflect our design intention of making them as references for projected visual vectors.

4.2 Comparison with Baselines

In this section, we will demonstrate the flexibility of PIKA by incorporating it with different backbones, including VGG and RESNET. We also investigate how significantly our proposed approach can improve the recognition accuracy compared to the baselines. The numerical results are presented in Table 2. As shown, PIKA, with all the backbones, enhances the performance by a large gap. Quantitatively, PIKA outperforms all compared SOTA significantly in terms of all metrics. Compared to VGG-16, PIKA improves the Precision, Recall, and *F1*-score by 5.36%, 11.44%, 9.49%, respectively. The performance gaps of PIKA to ResNet(s) are even more significant. Concerning all the settings of ResNet(s), PIKA improves the precision, recall, and *F1*-score by 45.83%, 58.67%, 58.11% on average, respectively. The most significant improvement can be found at ResNet-18, where PIKA improves the Precision, Recall, and *F1*-score by 48.41%, 68.18%, 66.20%, respectively.

Table 2. PIKA performace over different backbones. The best results are highlighted in **bold**.

Backbone		Precision	Recall	*F1*-score
VGG-16 [14]	Baseline	0.58967	0.47236	0.50121
	PIKA (ours)	**0.6213**	**0.5264**	**0.5488**
ResNet-18 [6]	Baseline	0.61020	0.49880	0.51570
	PIKA (ours)	**0.9056**	**0.8389**	**0.8571**
ResNet-34 [6]	Baseline	0.58200	0.49520	0.50870
	PIKA (ours)	**0.8832**	**0.8173**	**0.8315**
ResNet-50 [6]	Baseline	0.59612	0.51142	0.52146
	PIKA (ours)	**0.8664**	**0.7909**	**0.8101**
ResNet-101 [6]	Baseline	0.59120	0.50960	0.51620
	PIKA (ours)	**0.8148**	**0.7482**	**0.7609**

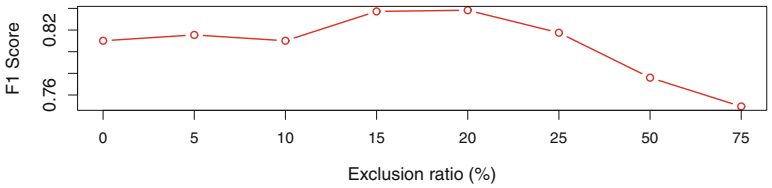

Fig. 5. *F1*-score of PIKA, given the Medical Knowledge Graph with different levels of edge cutting

4.3 Ablation Study

In this section, we first study the impacts of the Medical Knowledge Graph in Sect. 4.3. We then examine the impact of removing modules from the PIKA architecture on overall performance in Sect. 4.3 and 4.3.

Impacts of the Medical Knowledge Graph. When working with a graph, we should ensure that all the information from it is really beneficial for the model performance (containing no noise element). Since we built the Knowledge Graph by information from the set of prescriptions (Sect. 3.2), there are cases in which adding edges between some pills cause potential conflicts. We observe that while most edges have small weights, there are some with very large values. Those with small weights suggest they are potential noise that might hurt overall performance. With that in mind, we carry out an additional experiment for cutting edges and employ the *F1*-score for evaluation, which is plotted in Fig. 5. We first exclude 5% of edges with the lowest weights and increase the exclusion ratios up to 75%. The experiment results are shown in Fig. 5. As can be observed, some edges actually cause a negative impact on the overall result. The performance reaches its peak when excluding around 20% of edges with low weights and starts degrading afterward.

Table 3. PIKA performace with the Pseudo Classifier removal

Model	Precision	Recall	*F1*-score
ResNet-50 [6]	0.59612	0.51142	0.52146
PIKA	0.86640	0.79090	0.81010
PIKA Without Pseudo Classifier	0.67608	0.79778	0.69887

Impacts of the Pseudo Classification Layer. As declared in Subsect. 3.3, The Pseudo Classifier layer assists in removing redundant information from the MGK while retrieving condensed information about the pills in the input images. For this experiment, we use ResNet-50 as the backbone, and do training PIKA without Pseudo Classification Layer. The result is compared with the full version as well as our backbone. Specifically, employing the Pseudo Classifier improves the overall precision and *F1*-score by roughly 28% and 16%. The details of the result is presented in 3.

Impacts of the Projection Module, and Context Attention Module. Following, we study the performance of PIKA's when both Projection Moudule as well as Context Attention Module are removed. Instead of generating context vector c_i as the weighted sum of all condensed graph embeddings $q_i, i \in (0, \ldots, n)$, we directly take the mean over all q_i. For compliant with previous experiment, we also use ResNet-50 as our backbone as well as our baseline. As shown in Table 4, removing of the two modules leads to a degradation of 6% in the performance of PIKA.

Table 4. PIKA performace without Projection and Context Attention Modules.

Model	Precision	Recall	*F1*-score
ResNet-50 [6]	0.59612	0.51142	0.52146
PIKA	0.86640	0.79090	0.81010
PIKA w/o the Project and Attention Modules	0.82750	0.74700	0.76630

5 Conclusion and Future Work

In this study, we presented a novel approach to addressing challenges in image-based pill recognition. Specifically, we investigated a practical scenario aiming to identify pills from a patient's intake picture. The proposed method leverages additional information from prescriptions to improve pill recognition from photos. We first presented a method to construct a knowledge graph from prescriptions. We then designed a model to extract pills' relational information from the graph, and a framework to combine both the image-based visual and graph-based

relational features for identifying pills. Extensive experiments on our real-world pill image dataset showed that the proposed framework outperforms the baselines by a significant margin, ranging from 4.8% to 34.1% in terms of *F1*-score. We also analyzed the effects of the prescription-based medical knowledge graph on pill recognition performance and discovered that the graph's accuracy is critical in boosting the overall system's performance. We are actively developing this study by gathering more pill and prescription datasets required to verify the suggested technique and prove its usefulness in different clinical settings. We believe that leveraging the external knowledge will improve the accuracy of pill identification significantly.

Acknowledgement. This work was funded by Vingroup Joint Stock Company (Vingroup JSC), Vingroup, and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2021.DA00128.

References

1. World patient safety day (2022). https://www.who.int/news-room/events/detail/2022/09/17/default-calendar/world-patient-safety-day-2022. Accessed 14 Apr 2022
2. Chang, et al.: A deep learning-based intelligent medicine recognition system for chronic patients. IEEE Access **7**, 44441–44458 (2019). https://doi.org/10.1109/ACCESS.2019.2908843
3. Chang, et al.: Medglasses: a wearable smart-glasses-based drug pill recognition system using deep learning for visually impaired chronic patients. IEEE Access **8**, 17013–17024 (2020). https://doi.org/10.1109/ACCESS.2020.2967400
4. Chollet, et al.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, pp. 1251–1258 (2017)
5. Hang, J.Y., Zhang, M.L.: Collaborative learning of label semantics and deep label-specific features for multi-label classification. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 9860–9871 (2021). https://doi.org/10.1109/TPAMI.2021.3136592
6. He, et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
7. Hinton, et al.: Stochastic neighbor embedding. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems. vol. 15. MIT Press (2002). https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf
8. Li, Q., Qiao, M., Bian, W., Tao, D.: Conditional graphical lasso for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
9. Ling, et al.: Few-shot pill recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
10. Loshchilov, I., et al.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7

11. van der Maaten, L., et al.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008). http://jmlr.org/papers/v9/vandermaaten08a.html
12. Proma, et al.: Medicine recognition from colors and text. In: Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing. ICGSP 2019, pp. 39–43., Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3338472.3338484, https://doi.org/10.1145/3338472.3338484
13. Scott, et al.: Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd edn., March 2015. https://doi.org/10.1002/9781118575574
14. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 March 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1409.1556
15. Ting, H.W., et al.: A drug identification model developed using deep learning technologies: experience of a medical center in Taiwan. BMC Health Serv. Res. **20** (2020). https://doi.org/10.1186/s12913-020-05166-w, https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-020-05166-w#citeas
16. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , June 2016
17. Wang, Y., et al.: Multi-label classification with label graph superimposing. Proc. AAAI Conf. Artif. Intell. **34**(07), 12265–12272 (2020). https://doi.org/10.1609/aaai.v34i07.6909, https://ojs.aaai.org/index.php/AAAI/article/view/6909
18. Wang, Y., Xie, Y., Liu, Y., Zhou, K., Li, X.: Fast graph convolution network based multi-label image recognition via cross-modal fusion. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM 2020, pp. 1575–1584. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340531.3411880
19. Wong, Y.F., et al.: Development of fine-grained pill identification algorithm using deep convolutional network. J. Biomed. Inform. **74**, 130–136 (2017). https://doi.org/10.1016/j.jbi.2017.09.005, https://www.sciencedirect.com/science/article/pii/S1532046417302022
20. Yaniv, et al.: The national library of medicine pill image recognition challenge: an initial report. In: 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–9 (2016). https://doi.org/10.1109/AIPR.2016.8010584