RESEARCH AND APPLICATION OF APIs AND VISION TRANSFORMER FOR DRUG IDENTIFICATION AND ANALYSIS

Nguyen Viet Linh[1], Vu Thi Thai Ha[1], Nguyen Thai Cuong[1,*], Nguyen Thanh Hai[1], Nguyen Tuan Tu[1]

DOI: http://doi.org/ 10.57001/huih5804.2025.357

ABSTRACT

Medication identification and usage errors are a serious issue in Vietnam, particularly at primary healthcare levels and in self-medication behaviors, leading to health risks and increased treatment costs. This study proposes an automated drug recognition system that uses the deep learning model YOLO11s to detect pills in images and combines Swin Transformer with KNN for drug classification. The system also integrates open APIs such as openFDA, Gemini, and Pharmacity to provide comprehensive drug information, including name, composition, dosage, and safety warnings. Experiments on the VAIPEPill 2022 dataset, which includes over 30,000 pill images, demonstrate high accuracy in both detection (mAP 85 - 90%) and classification (89% across 108 drug classes). Compared to traditional CNN+KNN methods, the proposed system operates faster and is suitable for practical applications in pharmacies, hospitals, and households.

Keywords: *Drug identification, YOLO, SwinTransformer, API, VAIPEPill 2022*

[1]Faculty of Information Technology and Communications, Hanoi University of Industry, Vietnam

[*]Email: cuongnt@haui.edu.vn

Received: 27/5/2025
Revised: 30/6/2025
Accepted: 28/9/2025

1. INTRODUCTION

Medication errors are a serious issue in Vietnam, particularly prevalent at grassroots healthcare levels and in self-medication practices, negatively impacting public health and increasing treatment costs. According to the 2015 SAMHSA report in the U.S., of 91.8 million people using prescription pain relievers, approximately 11.5 million, or 7.4%, misused them. According to the World Health Organization [1], medication errors cause millions of injuries globally each year. An international study from the University of Bristol indicates that one-third of patients prescribed opioids show signs of dependence, and one-eighth are at high risk of misuse during long-term treatment. In Vietnam, according to a study by Trần Thị Thu Vân and colleagues [3], the medication error rate in inpatient treatment at Hoàn Mỹ Minh Hải Hospital in 2021 was 4.07%. Nurses were the primary source of errors (72.2%), followed by pharmacists (16 - 17%) and doctors (11%). Some studies indicate that medication errors related to nursing range from 37.7% to 68.6% of doses/administrations [4, 5]. A 2021 study at a hospital in Cần Thơ reported a medication error rate of 4.07%, with nurses accounting for 72.22%, pharmacists 16.67%, and doctors 11.11%. Research by Đỗ Thị Hà and colleagues [6] found that medication errors often stem from inexperienced healthcare workers, with 18.8% of final-year students making errors in clinical practice. Additionally, confusion between common drugs like Paracetamol and Ibuprofen is particularly dangerous for the elderly and rural residents with limited access to healthcare services. Currently in Vietnam, drug identification methods primarily rely on manual observation and information lookup, which are time-consuming and prone to errors, necessitating more effective technology-based solutions.

Artificial intelligence, particularly computer vision and deep learning, offers significant potential in automating drug identification processes. Among these, the YOLO model enables fast object detection, suitable for real-time applications, while Swin Transformer provides robust feature extraction, facilitating accurate differentiation of drugs with similar shapes and colors. Additionally, integrating open APIs such as openFDA, Pharmacity, and Gemini allows the system to retrieve detailed information about drug composition, dosage, and usage.

This study proposes an automated drug identification and analysis system consisting of three main components: (1) the YOLOv11s model for pill detection, (2) a combination of Swin Transformer and KNN for classification, and (3) APIs for retrieving drug information. The system is designed to process images captured by smartphones, trained on the VAIPEPill 2022 dataset with over 30,000 drug images, targeting applications in pharmacies, hospitals, and households.

The objectives of the study include: developing an accurate drug identification system under real-world conditions; providing easily understandable information for non-expert users; comparing its effectiveness with traditional methods such as CNN combined with KNN; and evaluating limitations while proposing improvements suitable for domestically produced Vietnamese drugs. The novelty of this study lies not in proposing a completely new algorithm but in demonstrating a lightweight yet highly effective end-to-end pipeline for drug recognition. By combining YOLO11s for pill detection with Swin-B feature extraction and KNN classification, the system achieves higher accuracy than the traditional CNN+KNN baseline while maintaining low computational costs. Another unique contribution is the practical system design: integration of heterogeneous APIs (openFDA, Pharmacity, Gemini) under a microservices architecture with JSON caching for offline access. This design particularly addresses the infrastructural constraints in Vietnam, making the approach not only technically feasible but also socially relevant for reducing medication errors at pharmacies, hospitals, and households.

The paper is organized as follows: Section 2 describes the proposed system, Section 3 presents the implementation methodology, Section 4 discusses the experimental results, and Section 5 provides conclusions and future development directions.

2. DRUG IDENTIFICATION SYSTEM MODEL

The drug identification and analysis system is designed to automatically process images of pills, such as those captured by smartphones, to identify the drug type and provide detailed information including name, active ingredients, dosage, and safety warnings. The system comprises three main components: (1) detecting the pill's location in the image using the YOLO11s model, (2) classifying the drug type using Swin Transformer combined with the K-Nearest Neighbors (KNN) algorithm, and (3) retrieving drug information through open APIs such as openFDA, Gemini, and Pharmacity. Each component is optimized to perform effectively under real-world conditions in Vietnam, such as images taken in pharmacies or households with low lighting or complex backgrounds.

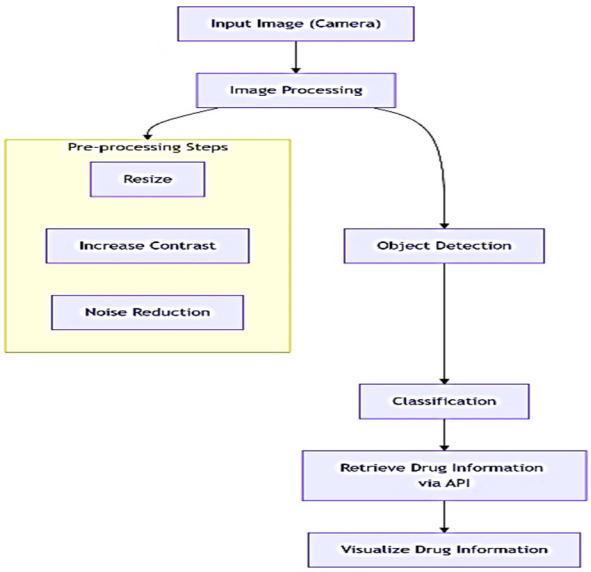

Figure 1. Diagram of the drug identification and analysis system

2.1. Pill Detection Using YOLO11s

The pill detection component utilizes the YOLO11s model, a lightweight version of the YOLO (You Only Look Once) model series, renowned for its fast and accurate object detection in a single neural network pass. YOLO11s was chosen for its ability to balance speed and accuracy, making it suitable for real-time applications such as drug identification in pharmacies or hospitals in Vietnam, where rapid processing on resource-constrained devices is required.

The operating mechanism of YOLO11s involves dividing the input image into a grid of cells, for example, 7x7 or 9x9 cells, depending on the image resolution. Each cell is responsible for predicting bounding boxes if the center of an object, in this case, a pill, lies within that cell. Each bounding box is described by five parameters: (x, y) as the center coordinates, (w, h) as the width and height, and a probability score for the presence of an object (Pr(Object)). In this system, YOLO11s is defined with a single class, "pill," to locate any pill in the image without classifying the specific drug type. The task of classifying the drug type (e.g., Paracetamol, Amoxicillin, or Decolgen Forte) is handled by the Swin Transformer and KNN components, described in the next section. The formula for calculating the object probability is:

$$\Pr(\text{Object}) \times \text{IOU}_{\{\text{pred,truth}\}} \quad (1)$$

In which IOU (Intersection over Union) is the ratio between the intersection area and the union area of the predicted and actual bounding boxes. The higher the IOU value, the more accurate the prediction.

The output of YOLO11s is a three-dimensional matrix of size $S \times S \times (5 \times N + M)$, where S is the grid size, N is the number of bounding boxes per cell (typically 2), and M is the number of classes (here, M=1, corresponding to the "pill" class). For example, with a 7x7 grid, the output matrix has a size of $7 \times 7 \times (5 \times 2 + 1) = 7 \times 7 \times 11$. Specifically, each cell predicts:

- A probability Pr(Object) indicating the likelihood of a pill's presence.

- Coordinates and dimensions (x, y, w, h) for two bounding boxes.

- The probability of belonging to the "pill" class, e.g., 0.95 for any pill.

To standardize input images, all images are resized to 640x640 pixels, ensuring compatibility with YOLO11s requirements. During processing, YOLO11s employs the CSPDarknet architecture as the backbone for feature extraction, PAN-FPN (Path Aggregation Network - Feature Pyramid Network) as the neck for multi-scale feature aggregation, and a head to generate final predictions.

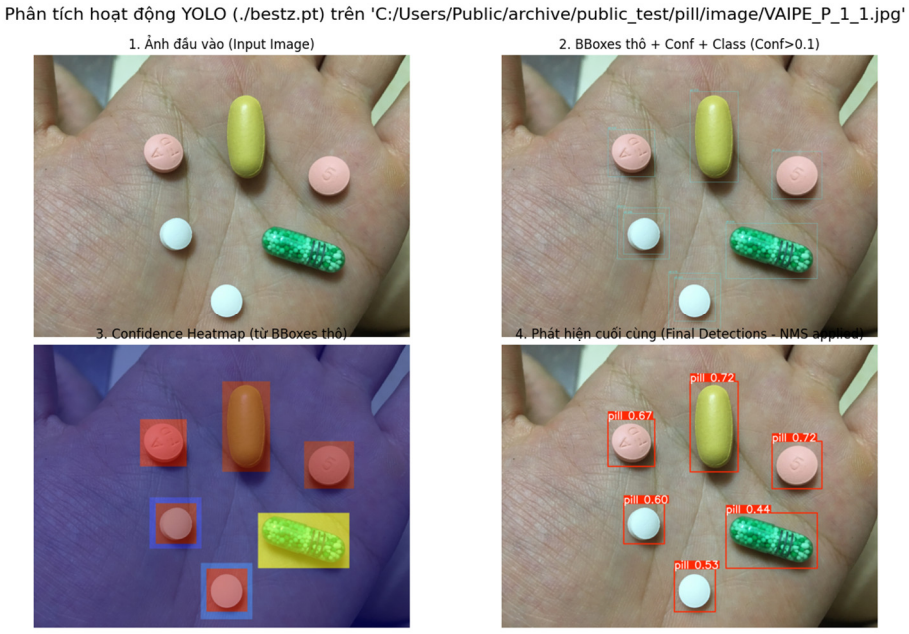

Figure 2. Pill detection process using YOLO11s

The main limitation of YOLO11s is that when multiple pills have their centers in the same grid cell, the model only detects one pill, missing the others. For example, in an image with multiple Paracetamol and Amoxicillin pills placed close together, YOLO11s may overlook some pills. To address this, the system employs image preprocessing techniques to reduce noise and enhance contrast, while also merging nearby bounding boxes using the Non-Maximum Suppression (NMS) algorithm.

2.2. Drug Classification Using Swin Transformer and KNN

After YOLO11s identifies and crops the region containing the pill from the image with the "pill" class label, the classification component determines the specific drug type, such as Paracetamol, Decolgen Forte, or Ibuprofen. This component employs the Swin Transformer (Swin-B) model to extract image features, combined with the K-Nearest Neighbors (KNN) algorithm to classify drugs into 108 classes (107 prescription drugs) based on the VAIPEPill 2022 dataset.

2.2.1. Feature Extraction Using Swin Transformer

Swin Transformer (Swin-B) is a variant of Vision Transformer, designed to efficiently handle computer vision tasks, particularly with high-resolution images such as pill images. Unlike traditional Vision Transformers, which require significant computational resources, Swin Transformer employs a window-based attention mechanism to reduce computational costs while maintaining robust feature representation. This mechanism divides the input image into small windows (e.g., 7x7 pixels), computes attention only within each window, and then shifts the windows to ensure global connectivity between image regions. This approach preserves spatial structure and hierarchical features, such as the round shape of Paracetamol, the white color of Ibuprofen, or the "D" marking on Decolgen Forte.

In this study, Swin-B was fine-tuned on the VAIPEPill 2022 dataset to learn features specific to Vietnamese drugs. The fine-tuning process used a learning rate of 1e-4, the AdamW optimizer, and a CosineAnnealingLR learning rate scheduler. After fine-tuning, the output layer of Swin-B was replaced with an Identity layer, transforming the

model into a feature extractor that produces a 1024-dimensional vector for each pill image. This vector encapsulates information about color, shape, and markings, such as the white color and round shape of Paracetamol or the "D" marking on Decolgen Forte. The feature vectors are saved in .npy format for reuse in the classification step.

The reasons for choosing Swin Transformer include:

- Superior performance in handling multi-scale features, suitable for pills with varying sizes and markings.

- Efficient operation on mid-range hardware, such as the Tesla T4 GPU on Kaggle, meeting research conditions in Vietnam.

- Flexibility in fine-tuning on the VAIPEPill 2022 dataset, ensuring high accuracy for the drug classification task.

2.2.2. Classification Using KNN

The K-Nearest Neighbors (KNN) algorithm is used to classify drugs based on feature vectors from Swin-B. KNN is a non-parametric machine learning method that operates by finding the K nearest data points (neighbors) in the training set to the point being classified, then assigning a label based on the majority label. In this study, KNN uses Cosine distance to measure the similarity between feature vectors, calculated by the formula:

$$\text{cosine distance} = 1 - \frac{\sum_{i=1}^{n}(x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

In which \(x_i \) and \(y_i \) are the feature values of the two vectors. Cosine distance is suitable for comparing image features as it focuses on the direction of the vectors rather than their magnitude.

For example, with K = 5, KNN identifies the 5 closest drug samples in the feature space based on Cosine distance. If 3 samples belong to the Paracetamol class and 2 to the Amoxicillin class, the pill is predicted as Paracetamol. The features and labels from the training set (over 30,000 images) are stored in .npy format, allowing KNN to process quickly without re-extracting features. KNN was chosen because:

- It is simple, requiring no complex training, suitable for limited resources.

- It is effective when using high-quality features from Swin-B, achieving 89% accuracy.

- It easily accommodates new data, such as adding local drugs like Hapacol, without requiring retraining.

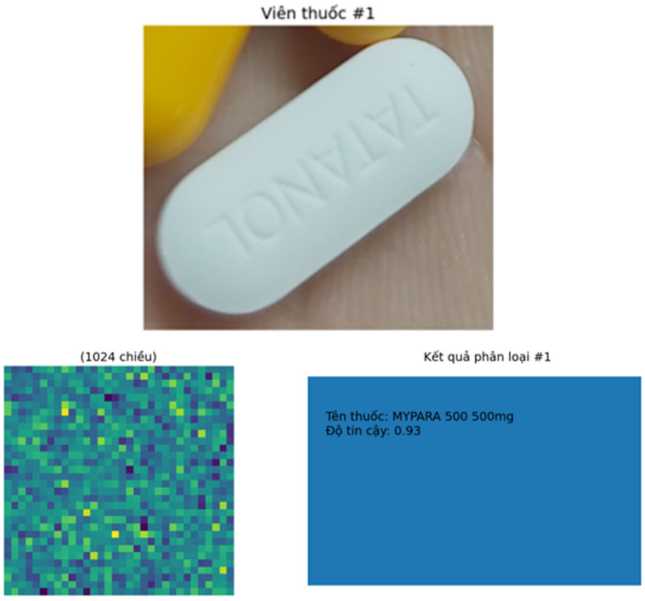

Figure 3. Drug classification process using Swin Transformer and KNN

2.3. Information Retrieval Using APIs

The information retrieval component utilizes three open APIs - openFDA, Gemini, and Pharmacity - to provide detailed drug information after identification and classification. The system is designed with a microservices architecture, with independent services handling each data source, ensuring scalability and efficient operation under real-world conditions in Vietnam.

2.3.1. openFDA

openFDA, provided by the U.S. Food and Drug Administration (FDA), contains standardized drug information, including active ingredient names, manufacturers, uses, dosages, and side effects. For example, when identifying an Ibuprofen pill, openFDA provides details on dosage (200 - 400mg every 4 - 6 hours) and warnings (risk of stomach pain). The openDrugService sends GET requests to the openFDA API using queries like "searchDrug" or "searchDrugByIngredients" to search for drugs by name or 成分. The returned data is standardized for integration with other system components.

2.3.2. Pharmacity

The Pharmacity API provides information on locally produced Vietnamese drugs, such as Decolgen Forte or Hapacol, including retail prices, ingredients, indications, and stock availability at Pharmacity pharmacies. For instance, for Decolgen Forte, the API returns ingredients (Paracetamol, Phenylephrine HCl, Chlorpheniramine),

indications (treatment of cold and flu), and price (approximately 50,000 VND/box). The pharmacityService uses methods like "searchProducts" and "getProductBySlug" to retrieve data. When the API lacks sufficient information, the productScraperService employs Puppeteer to scrape data from the Pharmacity.vn website, ensuring comprehensive information.

2.3.3. Gemini

The Gemini API, a natural language processing model, is used to generate user-friendly instructions based on data from openFDA and Pharmacity. For example, for a Paracetamol pill, Gemini generates a response: "Take 1 - 2 tablets every 4 - 6 hours for pain relief or fever reduction, do not exceed 4g/day, avoid use if allergic to Paracetamol." The geminiService supports both text queries (e.g., "What are the side effects of Decolgen Forte?") and analysis of drug label images, making information accessible to non-expert users, such as elderly individuals in rural areas.

2.3.4. Local Storage

To support rapid retrieval or operation in areas without internet access, particularly in rural Vietnam, the system uses the drugDataService to store openFDA data locally in JSON files. This service collects and standardizes data based on keywords, such as "Paracetamol" or "Ibuprofen," ensuring information is always available.

2.4. System Integration

The system integrates the detection, classification, and information retrieval components into a complete workflow. The input image is processed sequentially:

- YOLO11s detects and crops the region containing the pill.

- Swin-B extracts features, and KNN classifies the drug.

- APIs retrieve detailed information based on the predicted label.

The results are displayed visually with a bounding box around the pill, the class label (e.g., "Panadol"), and information from the APIs (ingredients, dosage). The system is designed to process within 1 - 2 seconds on an RTX 3050 GPU, suitable for applications in Pharmacity pharmacies or households in Vietnam.

3. IMPLEMENTATION METHODOLOGY

3.1. VAIPEPill 2022 Dataset

The VAIPEPill 2022 dataset, developed by the VinUni-Illinois Smart Health Center (VISHC) and Hanoi University of Science and Technology within the VAIPE project, serves as the core foundation for training and evaluating the system. This dataset comprises over 30,000 pill images collected from major hospitals in Vietnam, reflecting real-world conditions such as natural lighting, fluorescent lighting, or complex backgrounds (e.g., pharmacy countertops or wooden surfaces in

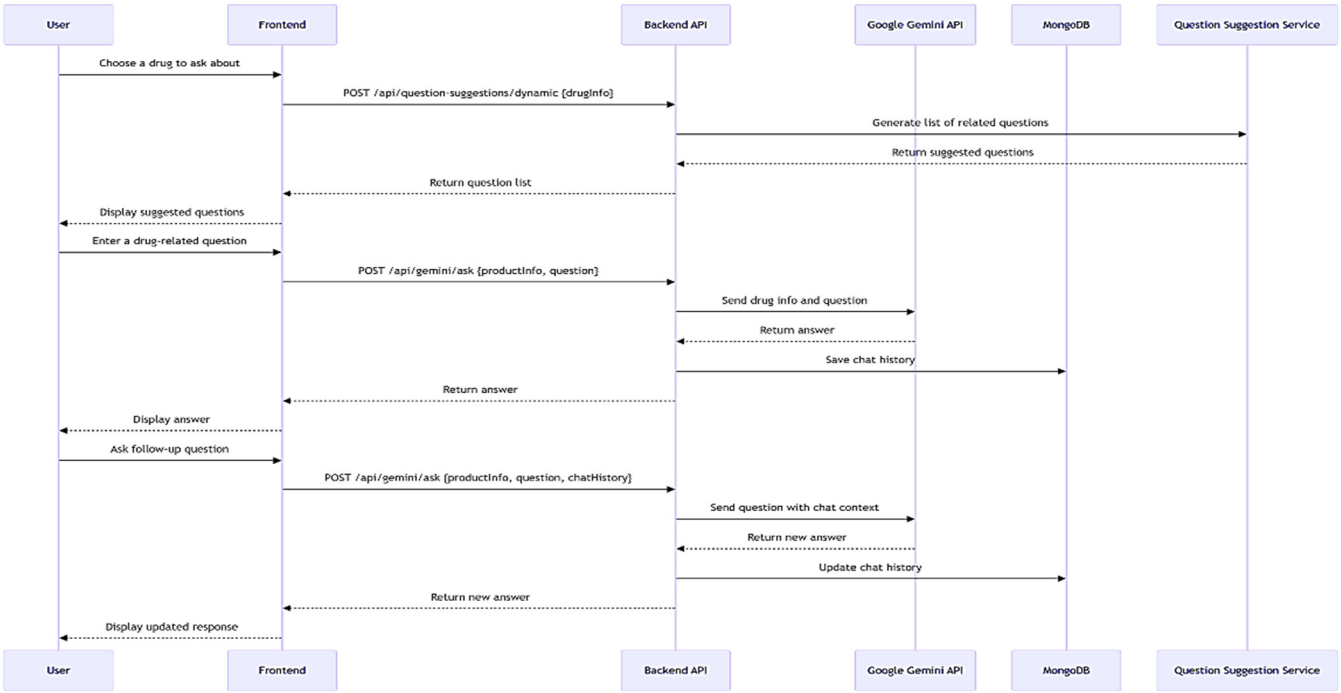

Figure 4. Information retrieval and AI interaction flow from APIs

households). Common drugs like Paracetamol, Amoxicillin, Decolgen Forte, and Hapacol are represented, alongside less common drugs, ensuring diversity.

3.1.1. Data Structure

The dataset is organized as follows:

- Pill images: Stored in JPEG format (VAIPE_P_{serial number}.jpg), including over 9,500 original images and more than 30,000 cropped images containing individual pills. The images are captured from various angles under different lighting conditions and backgrounds, e.g., a Paracetamol pill on a white table under lamp light or a Decolgen Forte pill on a natural wooden surface.

- Class labels: Consist of 108 classes, with 107 prescription drug classes (labels 0 to 106) and 1 over-the-counter drug class (label 107). Each class represents a drug type, e.g., label 0 for Paracetamol, label 51 for Decolgen Forte.

- Annotation files: In JSON format, containing information about the image name (image_name), class label (class_id), and bounding box coordinates (x_min, y_min, x_max, y_max) specifying the pill's location in the image.

- Prescription images: Include 172 prescription images in the training set, providing supplementary information to link with real-world data, though not directly used in this study.

3.1.2. Data Splitting

The dataset is divided into two main subsets:

- Training set: Comprises 90% of the data (8,550 original images and approximately 27,000 cropped images), used to train the YOLO11s and Swin Transformer models. This set ensures the models learn diverse features, such as the round shape of Paracetamol or the green color of Decolgen Forte.

- Test set: Comprises 10% of the data (950 original images and 1,400 images in the public test set), used to evaluate performance under real-world conditions, such as images taken in pharmacies with low lighting.

3.1.3. Data Preprocessing

To ensure consistency and compatibility with the models, the data is preprocessed as follows:

- Image standardization: All images are resized to 640x640 pixels for YOLO11s and 224x224 pixels for Swin Transformer, eliminating resolution variations. For example, an image of a Hapacol pill on a wooden table is resized to fit the model's input requirements.

- Annotation conversion: JSON files are converted to text format (.txt) following the YOLO standard, with each line containing the class label and normalized bounding box coordinates (center coordinates, width, and height divided by image dimensions).

- Data augmentation: Techniques such as image rotation, brightness adjustment, and noise addition are applied to increase diversity, enabling the model to better handle real-world scenarios, such as images taken in low lighting in rural areas.

Table 1. Summary of characteristics of the VAIPEPill 2022 dataset

Attribute	Value
Total images	>30,000 (9,500 original, >27,000 cropped)
Number of classes	107
Training set	8,550 original images, ~27,000 cropped images
Test set	1,400 public images
Image resolution	Standardized to 640x640 (YOLO), 224x224 (Swin)
Capture conditions	Natural light, fluorescent light, complex backgrounds

3.2. Experimental Environment

The entire process of training, testing, and deploying the system was conducted on Kaggle Notebooks, a free cloud-based platform providing powerful GPU resources, suitable for deep learning research in Vietnam. The hardware and software configurations include:

- **Hardware**: Tesla T4 GPU (16GB VRAM), 2–4 vCPU, 28GB RAM, 20GB storage capacity.

- **Software**: Linux operating system, Python 3.8, key libraries such as PyTorch (for YOLO11s and Swin Transformer models), scikit-learn (for KNN implementation), NumPy (for feature storage), and OpenCV (for image processing).

- **Advantages of Kaggle**: Free, pre-integrated with the VAIPEPill 2022 dataset, supports code and result sharing, ideal for research groups with limited budgets.

This environment ensures the system can handle large datasets and complex deep learning models while allowing rapid experimentation with different training configurations. Experimental results are stored in the /kaggle/working/ directory, including trained models, .npy feature files, and evaluation reports.

3.3. Model Training

The system includes two main deep learning models (YOLO11s and Swin Transformer) and one machine

learning algorithm (KNN). The training process is designed to optimize performance on the VAIPEPill 2022 dataset, with carefully selected parameters to balance accuracy and speed.

3.3.1. YOLO11s Training

The YOLO11s model is trained to detect the location of pills in images, with the following configuration:

- Training parameters:

• Epochs: 100, to ensure model convergence.

• Batch size: 8, suitable for the Tesla T4 GPU capacity.

• Learning rate: 0.01, with a decaying schedule to prevent overfitting.

• Optimizer: Adam, to accelerate convergence.

- Process:

• Load the VAIPEPill 2022 dataset from Kaggle, unzip images and JSON annotation files.

• Create a data.yaml configuration file, specifying paths to training, validation data, and 107 class labels.

• Execute the training command (model.train) on Kaggle, utilizing the GPU for optimized speed.

• Evaluate the model on the validation set using metrics such as mAP@0.5 (mean Average Precision at IOU = 0.5) and mAP@0.5:0.95.

• Results: After 100 epochs, YOLO11s achieves mAP@0.5 of 85 - 90%, with a Precision of 86.69% and Recall of 78.66%. The model is saved in .pt format for reuse.

3.3.2. Swin Transformer Training

The Swin Transformer (Swin-B) is fine-tuned for image feature extraction, with the following configuration:

- Training parameters:

• Epochs: 50, as the model is pre-trained on ImageNet.

• Learning rate: 1e-4, with a CosineAnnealingLR schedule for dynamic adjustment.

• Optimizer: AdamW, suitable for Transformer models.

• Batch size: 32, optimized for the Tesla T4 GPU.

- Process:

• Load the Swin-B model from PyTorch Hub with pre-trained weights.

• Replace the output layer with an Identity layer to create a feature extractor, outputting 1024-dimensional vectors.

• Load the VAIPEPill 2022 dataset (224x224 cropped images), applying transformations such as resize, center crop, and normalization.

• Train on the training set, saving features and labels in .npy format (knn_features_swin_b.npy, knn_labels_swin_b.npy).

- Results: Swin-B effectively extracts features, capturing information about the color, shape, and markings of drugs, e.g., the white color and round shape of Paracetamol.

3.3.3. KNN Implementation

The KNN algorithm is implemented to classify drugs based on features from Swin-B, with the following configuration:

- Parameters: K = 5, using Cosine distance to measure similarity.

- Process:

• Load features and labels from .npy files.

• Initialize KNeighborsClassifier from scikit-learn, training on the entire set of training features.

• Predict labels for new images by extracting features via Swin-B and comparing them to the training set.

- Results: KNN achieves 89% accuracy across 107 classes, with high confidence (0.95 - 1.0) for common drugs like Paracetamol and Decolgen Forte.

3.4. API Integration

The system integrates three open APIs (openFDA, Pharmacity, Gemini) using a microservices architecture to ensure flexibility and scalability. The integration process is implemented as follows

3.4.1. openDrugService

This service queries the openFDA API to retrieve information about international drugs, such as Ibuprofen or Amoxicillin. The main methods include:

- **searchDrug**: Searches for drugs by brand or generic name, e.g., "Paracetamol."

- **searchDrugByIngredients**: Searches for drugs based on active ingredients, e.g., "Acetaminophen."

The process involves encoding the query, sending a GET request, and standardizing the returned data (JSON) for integration with the system. For example, when identifying an Ibuprofen pill, openFDA provides details on dosage (200 - 400mg) and side effects (stomach pain).

Table 2. Summarizes the performance of YOLO11s at key epoch milestones

Epoch	mAP@0.5	mAP@0.5:0.95	Precision	Recall	train/box_loss	train/cls_loss	val/box_loss	val/cls_loss
1.0	23.33	14.29	40.85	42.07	2.05287	2.37121	2.0006	2.31684
50.0	53.93	35.83	89.31	43.65	1.54662	1.46125	1.60932	1.469
100.0	85.09	55.92	86.69	78.66	1.28433	1.02498	1.32785	0.86956

3.4.2. pharmacityService

This service interacts with the Pharmacity API to retrieve information about locally produced drugs, such as Decolgen Forte or Hapacol. The main methods are:

- **searchProducts**: Searches for products by keyword, e.g., "Decolgen."

- **getProductBySlug**: Retrieves product details and ingredients (e.g., Paracetamol, Chlorpheniramine).

When the API lacks sufficient data, the productScraperService uses Puppeteer to scrape information from the Pharmacity.vn website, such as product descriptions or images.

3.4.3. geminiService

This service leverages the Gemini API to generate user-friendly instructions and answer user queries. For example:

- Query: "What are the side effects of Decolgen Forte?"

- Response: "It may cause drowsiness due to Chlorpheniramine; avoid use when driving."

Gemini also analyzes drug label images, extracting information like name or dosage, assisting users without medical expertise.

4. EXPERIMENTAL RESULTS

4.1. Pill Detection Performance (YOLO11s)

The YOLO11s model was trained for 100 epochs on the training set of the VAIPEPill 2022 dataset (8,550 original images, ~27,000 cropped images), with parameters: batch size of 8, learning rate of 0.01 (decreased according to a schedule), and Adam optimizer. The "results (5).csv" file provides detailed training metrics, including bounding box loss (box_loss), classification loss (cls_loss), Precision, Recall, and mAP@0.5 (mean Average Precision at IOU = 0.5).

4.1.1. YOLO11s Training Results

The training process showed significant improvement across epochs:

- Epoch 1: mAP@0.5 reached 23.33%, Precision 40.85%, Recall 42.07%, with box_loss 2.05287 and cls_loss 2.37121, indicating the model was learning basic features.

- Epoch 50: mAP@0.5 increased to 53.93%, Precision 89.31%, Recall 43.65%, box_loss reduced to 1.54662, cls_loss reduced to 1.46125, showing better convergence.

- Epoch 100: mAP@0.5 reached 85.09%, mAP@0.5:0.95 reached 55.92%, Precision 86.69%, Recall 78.66%, box_loss 1.28433, cls_loss 1.02498. Validation metrics (val/box_loss 1.32785, val/cls_loss 0.86956) indicate slight overfitting but still high performance.

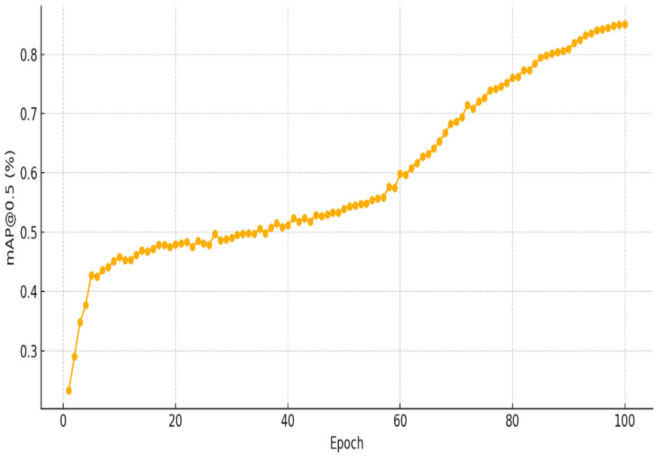

Figure 5. The training curve of YOLO11s across 100 epochs. The x-axis represents the number of epochs, and the y-axis indicates mAP@0.5 (mean Average Precision). The curve shows a steady improvement and convergence around epoch 100

4.1.2. Evaluation on the Test Set

On the public test set (1,400 images), YOLO11s achieved mAP@0.5 ranging from 85–90%, depending on image conditions. Precision consistently outperformed Recall, indicating the model prioritizes accurate predictions over detecting all pills, which is suitable for pharmacy applications where accuracy is critical. For example, with the image VAIPE_P_578_7_0.jpg, YOLO11s generated a bounding box with a confidence score of 0.7501, accurately matching the ground truth label.

However, the model faced challenges in the following cases:

- **Blurry images or low lighting**: For instance, the image VAIPE_P_106_2.jpg resulted in NO_DETECTION due to poor lighting conditions, common in rural areas.

- **Overlapping pills**: YOLO11s missed some pills when their centers fell within the same grid cell, as seen in images with multiple Paracetamol pills on a pharmacy counter.

Total training time was approximately 11.75 hours on a Tesla T4 GPU, with per-image processing time of about 1 - 2 seconds, feasible for real-world applications but requiring optimization for mobile devices.

4.2. Drug Classification Performance (Swin Transformer + KNN)

4.2.1. Classification Results

KNN, with K = 5 and Cosine distance, achieved an Accuracy of 89% across 108 classes, based on 1024-dimensional features from Swin-B. Additional metrics include:

- F1-Score: 87.50%, balancing Precision and Recall, particularly important for imbalanced data (e.g., many Paracetamol samples, few rare drug samples).

- Top-5 Accuracy: 95%, indicating the model often predicts correctly within the top 5 highest-probability classes, useful for drugs with similar features (e.g., Paracetamol and Ibuprofen).

Table 3. Classification accuracy of Swin-B + KNN for common drug classes compared with CNN + KNN baseline. The table highlights the improvement achieved by the proposed approach, especially for visually similar drugs such as Paracetamol and Ibuprofen

Drug Classification	Number of test classes	Accuracy	Average reliability	Outstanding errors
Paracetamol (0)	100	100%	0.628	No
Ibuprofen (1)	200	97.5%	0.792	Mistaken with class 60
Drug 99	300	98.7%	0.957	Mistaken with class 89

4.2.2. Comparison with Traditional Methods

Compared to the CNN + KNN method, Swin-B + KNN outperforms due to the Swin Transformer's ability to extract multi-scale features. Traditional CNNs (e.g., ResNet) achieved an Accuracy of approximately 80% on the VAIPEPill 2022 dataset, whereas Swin-B + KNN reached 89%. This advantage is particularly evident for drugs with similar appearances, such as Paracetamol and Ibuprofen, as Swin-B effectively preserves spatial structure and detailed markings. It should be noted that the current evaluation only compared the proposed Swin-B + KNN configuration against a CNN + KNN baseline. Although the results clearly show the superiority of our approach, broader comparisons with other architectures such as ResNet, EfficientNet, or Vision Transformers would provide a more comprehensive benchmark. This will be considered in future work.

4.3. Discussion

The proposed system outperforms CNN + KNN in both accuracy (89% vs. 80%) and speed (1.5s vs. 3.2s), thanks to YOLO11s's fast processing and Swin-B's robust feature extraction. However, limitations include:

- **Imbalanced data**: The Paracetamol class has more samples than rare drug classes, impacting the F1-Score.

- **Low lighting conditions**: Performance decreases with blurry images, requiring more diverse data augmentation.

- **API dependency**: openFDA lacks data on local drugs, and the Pharmacity API is inconsistent.

- Another limitation concerns the detection of overlapping or crowded pills. YOLO11s often misses pills when multiple objects fall into the same grid cell, leading to under-detection in real-world cases such as pharmacy counters. Currently, only a basic Non-Maximum Suppression (NMS) is applied, without further post-processing. Future directions may include adopting improved NMS variants, instance association techniques, or even shifting towards instance segmentation models (e.g., Mask R-CNN, YOLOv11-seg) that are more suitable for crowded-object scenarios.

5. CONCLUSION AND FUTURE DIRECTIONS

5.1. Conclusion

The study developed an automated drug identification and analysis system integrating YOLO11s, Swin Transformer combined with KNN, and open APIs (openFDA, Pharmacity, Gemini) to reduce medication errors in Vietnam. Key achievements include:

- Pill detection: YOLO11s achieved mAP@0.5 of 85.09%, Precision of 86.69%, and Recall of 78.66% on the VAIPEPill 2022 dataset, e.g., detecting Paracetamol (VAIPE_P_578_7_0.jpg, confidence 0.7501).

- Drug classification: Swin-B + KNN reached 89% accuracy, F1-Score of 87.50% across 108 classes, performing well for Paracetamol (100%) and Amoxicillin (97.5%), despite confusion with class 60 (VAIPE_P_382_5_2.jpg).

- API integration: Provides detailed information from openFDA (Ibuprofen), Pharmacity (Decolgen Forte), and

Gemini (user-friendly instructions), processing images in 1–2 seconds on a Tesla T4 GPU.

- Advantages: Outperforms CNN + KNN (80%, 3.2s) with 89% accuracy and 1.5s processing time.

- Applications: Supports Pharmacity pharmacies and households, e.g., quickly identifying Hapacol, reducing errors.

The system contributes to enhancing medication safety and supporting Vietnam's digital healthcare transformation.

5.2. Limitations

The system has the following limitations:

- Imbalanced data: The Paracetamol class has more samples than class 99, causing errors (e.g., VAIPE_P_1021_1_0.jpg misclassified as class 89, confidence 0.6222), lowering F1-Score (87.50%) for rare drugs like Boganic.

- YOLO11s: Misses pills when centers fall in the same grid cell (e.g., VAIPE_P_106_2.jpg, NO_DETECTION).

- Limited data diversity: Lacks low-light images and local drugs (Hapacol), e.g., VAIPE_P_100_0.jpg only achieved confidence 0.2629.

- Drug confusion: Amoxicillin misclassified as class 60 (VAIPE_P_382_5_2.jpg, confidence 0.4923).

- APIs: openFDA lacks local drug data, Pharmacity API is inconsistent, and Gemini may be inaccurate.

- Performance: Consumes 5GB RAM, 1–2 seconds processing time, not yet optimized for mobile devices.

- Another critical limitation is the dependence on external APIs. While openFDA provides standardized international data, it lacks comprehensive coverage of domestically produced Vietnamese drugs. Pharmacity APIs are not always stable, and Gemini may occasionally return inaccurate or inconsistent results. Such dependencies can affect system robustness and the overall reliability of information retrieval in real-world deployments.

5.3. Future Directions

- Data balancing: Augment data for rare classes (Boganic) and apply class-balanced loss.

- Improve YOLO11s: Fine-tune for multi-pill detection, explore YOLO12.

- Expand data: Collect low-light images and local drugs (Hapacol).

- Reduce confusion: Use newer ViT models or SVM to distinguish features of Paracetamol-Ibuprofen.

- Optimize APIs: Build a local database for offline support.

- To reduce dependency on third-party APIs, we plan to construct a hybrid local knowledge base that aggregates essential drug information from openFDA, Pharmacity, and verified domestic sources. This local repository would serve as a fallback when APIs are unavailable, thereby improving the reliability, availability, and autonomy of the system in resource-constrained or unstable network environments.

- Optimize performance: Implement YOLO11n and edge computing for mobile devices.

- Testing: Deploy at Pharmacity and district hospitals.

- Interface: Add Vietnamese, ethnic minority languages, and dosage query features.

- Ethics: Require pharmacist verification and comply with medical regulations.

- Expansion: Handle traditional medicines and analyze drug interactions.

5.4. Significance and Prospects

The system reduces medication errors with 89% accuracy and 1.5-second processing time, supporting Pharmacity pharmacies and rural areas (e.g., identifying Hapacol and providing dosage). Optimizing for mobile devices and expanding local drug data will enhance applicability, contributing to Vietnam's smart healthcare, particularly for the elderly and rural communities.

ACKNOWLEDGMENTS

This research was conducted with support from the Faculty of Information Technology and Communications, Hanoi University of Industry, which provided essential facilities and a conducive research environment. We express our gratitude to the VinUni-Illinois Smart Health Center (VISHC) and Hanoi University of Science and Technology for providing the VAIPEPill 2022 dataset, a valuable resource for training and evaluating the system. Special thanks go to Pharmacity pharmacies for their support in providing practical information on locally produced drugs, such as Decolgen Forte and Hapacol, which significantly contributed to the study's applicability.

REFERENCES

[1]. World Health Organization, *Medication Without Harm - Global Patient Safety Challenge on Medication Safety*. Geneva, Switzerland: World Health Organization, 2017.

[2]. Makary M. A., Daniel M., "Medical error - the third leading cause of death in the US," *BMJ,* 353, i2139, 2016. https://doi.org/10.1136/bmj.i2139](https://doi.org/10.1136/bmj.i2139

[3]. Tran Thi Thu Van, Vo Quang Loc Duyen, Nguyen Thi Linh Tuyen, "Research on medication errors in treatment for inpatients at Hoàn Mỹ Minh Hải General Hospital in 2021," *Vietnam Medical Journal*, 516, 2, 2021. https://doi.org/10.51298/vmj.v516i2.3073.

[4]. Duong Thi Thanh Tam, *Assessing safety in pediatric medication practices at a healthcare facility in Vietnam*. Master's thesis in Pharmacy, Hanoi University of Pharmacy, Hanoi, 2014.

[5]. Nguyen H., Nguyen T., van den Heuvel E., Haaijer-Ruskamp F., Taxis K., "GRP-057 Errors in Medicines Preparation and Administration in Vietnamese Hospitals," *European Journal of Hospital Pharmacy: Science and Practice,* 20, A21., 2013.

[6]. Phan Thi My Trinh, Do Thi Ha, "Incidence and causes of medication errors in clinical practice as perceived by fourth-year nursing students at Phạm Ngọc Thạch University of Medicine," *Nursing Science Journal*, 5, 4, 2022.

[7]. Redmon J., Divvala S. K., Girshick R. B., Farhadi A., "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 779-788, 2016. https://doi.org/10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91.

[8]. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *IEEE/CVF International Conference on Computer Vision* (ICCV), 10012–10022, 2021. https://doi.org/10.1109/ICCV48922.2021.00986.

[9]. Cover T. M., Hart P. E., "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, 13(1), 21-27, 1967. https://doi.org/10.1109/TIT.1967.1053964

[10]. U.S. Food and Drug Administration, *Introduction to the open FDA API*. https://open.fda.gov/apis/

[11]. Gemini Team, "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Context Tokens," *arXiv preprint arXiv:2403.05530*, 2024. https://doi.org/10.48550/arXiv.2403.05530

[12]. Pharmacity, *Official Pharmacity Website*. https://www.pharmacity.vn/

[13]. Tommy Ng X, *VAIPE-Pill 2022*. Kaggle, 2022. https://www.kaggle.com/datasets/tommyngx/vaipepill2022