P.I.L.L.: A Deep Learning Model for Prescription Pill Recognition and Classification

Jett Badalament-Tirrell
School of Data Science
University of Virginia
hwq8mr@virginia.edu

Olivia Byram
School of Data Science
University of Virginia
ocb3wv@virginia.edu

Hamsini Muralikrishnan
School of Data Science
University of Virginia
hm7qgr@virginia.edu

I. Motivation

It is estimated that ∼68% of all Americans take at least one daily prescription pill, and a large amount of that percentage take more than one. It is common practice to put daily prescriptions and vitamins in pill cases, whether it be for easy storage, travel, or simply to remember if they have taken their daily pills. However, many of those pills look incredibly similar, and it is all too easy to mix them up when they reside in the same section of a pill case. This can lead to confusion, and often results in someone taking the wrong medication at the wrong time. While this is not typically a big deal, this could be life threatening for some people on necessary medications. The FDA requires that all pills have a distinct shape, color, size, score, and imprint code. However, the scoring and imprints tend to rub off and there are only so many colors, sizes, and shapes that a pill can be. We propose a solution to this being a deep learning model that is able to take an input of an image of a pill, then output the name of the pill. The name of this model is P.I.L.L. (Pill Identification and Labeling Locator). Our research question is 'Can we develop a deep-learning model that accurately identifies and classifies prescription pills from images using visual and textual features?'

II. Dataset

The dataset that we will be using for our project P.I.L.L. is a "Pharmaceutical Tablets Dataset" from Kaggle [4] (public URL is https://www.kaggle.com/datasets/trumedicines/pharmaceutical-tablets-dataset). This is an open-source dataset and contains 252 unique pill images, where each pill image has been convoluted to create 20,000 images by rotating the pills, adding a gray scale, etc. The unique pill images vary in pill size, color, shape, and imprint, allowing a dataset that is diverse. By using a dataset that includes multiple angles of the same pill types, we are training the model to handle images that are not perfectly photographed. We will be doing a 60/40 training test split with the model, consistent with the studies that we researched in our literature review.

III. Literature Review

A. An Accurate Deep Learning Based System For Automatic Pill Identification: Model Development and Validation

The paper 'An Accurate Deep Learning Based System For Automatic Pill Identification: Model Development and Validation' [1] covers the process to build a deep learning model for pill identification. In this study, there were 2 steps in the pill identification process: pill recognition and pill retrieval for accurate outputs. For pill recognition, the model focuses on key attributes such as color, shape form, and the embossing on the pill itself. These are all unique features to each pill, which is what would allow a model to accurately learn nuances of different pills to predict their names. Pill retrieval is then done using those features to retrieve the pill name from a database. In the pill recognition process, the attributes of the pills are used to classify the pill image itself along with passing these attributes/pill name into an LLM to improve the accuracy and performance. For the training and test data, a ∼40% to ∼60% split was used. To effectively test accuracy of the model, the validation/test set consisted of all pill images that the model had not seen or been trained on before. For each pill, there was a photo of the front and back of the pill included in the training/test sets.

For the pill recognition step, the image is first passed through YOLOv5, an object detection model and ResNet 32, an image classification model. YOLOv5 detects the pill, extracting the embossed characters from the pill and tracking their coordinates. This information is passed to ResNet32, which recognizes and classifies the tablet's attributes. These characteristics are passed through a character language model which uses the information from the YOLOv5 and ResNet 32 models to correct the embossed characters on the pill, putting them in order. Finally, the model transitions to the pill retrieval step where similarity metrics are calculated and used to rank the images against their labels. This allows the model to identify the pill image in the final output.

The dataset included 24,404 pill images from both South Korea and the United States. The model exhibited ∼85.6% accuracy on pills from South Korea and ∼74.5% on pills from the United States. As part of this work, the researchers were able to confirm that the model is able to accurately recognize and identify new pills without model updates or

excessive fine tuning, making it computationally efficient with high performance. Finally, an ablation study was used to determine whether the character language model was needed. The ablation study found it to be a crucial step in improving accuracy in the pill identification step.

While reading this work, we identified gaps that we questioned and plan to resolve within our own work. One gap was whether the model is built to handle images that a patient/consumer takes of the pill. Many training/validation data images come from established databases where the image has been taken in good lighting against a high contrast background to provide the best image quality. Additionally, this paper takes images that have pills oriented "properly", and each input image includes the front and back of the pill. This may be a limitation if a consumer/patient is trying to use this model for quick identification of their medicine. To resolve this, our dataset includes images of pills in different orientations so the model isn't sensitive to the pill being "right side up" for every input image. We hope this helps our model to effectively filter out noise.

Overall, this paper was helpful to lay the foundation on what steps we will need to answer our research question. We will use aspects of this work as a starting point for models to leverage and this study's metrics as a gauge on our model performance.

B. Real-time pill identification and classification using deep learning framework for medicine inspection systems

The next study that we identified was 'Real-time pill identification and classification using deep learning framework for medicine inspection systems' [2]. This study used deep-learning for prescription pill identification and imprint recognition. It followed a similar methodology to the paper above, although its imprint regions were processed in a different manner.

The study began by using YOLOv5 to localize pharmaceutical pills and their surface imprints within natural RGB image. YOLOv5 is able to identify pills and their imprints, allowing for the detected images to be put into a ResNet-32 feature extractor. This extracts shape, color, size, and other useful physical characteristics. While this is happening, the images are also being passed through what is called a 'Deep Text Spotter' model. This model allows for pictures that are not professional, meaning they could have poor lighting, uneven orientation, blurriness, or even some sort of obstruction over a small part of the pill. This allows the model to account for pictures that are taken at home by normal users of this deep-learning model. Finally, the output is refined by a recurrent Neural Network, allowing for corrections on any significant character errors.

This paper found that the proposed framework achieved a detection accuracy of $\sim$97.8% outperforming baseline models across precision, recall, and F1-score metrics. While this is a significantly high score, we identified a few gaps that we hope to address within our own methodology. One thing that we recognized was that the use of the 'Deep Text Spotter' (DTS) model may not fully resolve the issue of poor images. DTS has proven somewhat helpful, but there are other models that could be more beneficial with our training/testing dataset. For that reason, we will be using ASTER as well and comparing the results. Additionally, the dataset used in this study did not include many wrongly-oriented pill pictures, meaning that it may not be trained well enough to handle them. Our dataset includes thousands of pills that are not oriented correctly to resolve this gap in research. We are unsure of how diverse the dataset in this study was, which could be a potential gap. Our dataset also includes a diverse subset of prescription pills, many of which have very subtle differences to help our model generalize more accurately.

This study was incredibly helpful in pointing out different avenues to explore for our proposed method, and we hope to achieve similar results.

C. High Accurate and Explainable Multi-Pill Detection Framework with Graph Neural Network-Assisted Multimodal Data Fusion

Another relevant study in this space is "High Accurate and Explainable Multi-Pill Detection Framework with Graph Neural Network-Assisted Multimodal Data Fusion" [3]. This research tackled what the researchers saw as a gap present in many prior works, which is the assumption that a pill identification task would involve only a single pill captured under ideal lighting and background conditions. Instead, Nguyen and colleagues developed a model capable of handling complex scenarios where multiple pills appear together, sometimes overlapping or even blending in together. This shift toward recognizing multiple pills might make their model more applicable to cases such as home medication or pharmacy automation.

To achieve this, [3] proposed a two-stage framework that first detects pills then models their relationships. This combined a convolutional neural network backbone for feature extraction with a Graph Neural Network or GNN for relational reasoning. After detecting each pill within an image, the system built a graph where nodes represented individual pills and edges encoded contextual relationships like if the pills occurred together more often, spatial proximity, and color or shape similarity. This graph structure allowed the model to interpret the relationships between pills jointly, instead of treating each one as an isolated object. The integration of multimodal data resulted in stronger performance compared to baselines such as Faster R-CNN and YOLOv5, and also made the system more explainable, which could be especially useful if the model were to be used and updated based on user feedback and labeling, as it could then give insight into which pills are seen together more often in real scenarios.

Although [3] achieved impressive results, we noted a few gaps that guide our project's direction. First off, we feel that in practice it is more likely that individuals would be trying to identify one pill at a time, and if the model expects multiple pills then we expect it to perform a bit worse in that case. The dataset used in their study, while more diverse than in

previous work, still lacked variation in lighting, camera angle, and pill wear, which are factors that can significantly impact real-world performance. Additionally, their model relied on predefined graph relationships, which may not generalize well to unseen pill classes or newly introduced medications. These limitations informed our decision to train primarily on single-pill detection under variable real-world imaging conditions as opposed to multi-pill arrangements. We plan to focus on improving robustness to image quality and pill orientation, and to explore adaptive approaches that can dynamically learn new relationships from data rather than relying on static graph structures.

IV. Proposed Method

The initial approach that our group will be taking for our project closely resembles the approaches that were used in the papers 'An Accurate Deep Learning Based System For Automatic Pill Identification: Model Development and Validation' [1] and 'Real-time pill identification and classification using deep learning framework for medicine inspection systems' [2]. We propose a method that integrates key aspects of the methodology of both studies, while addressing the gaps in research that we identified within them.

As all three studies in our literature review did, we will focus our model on two primary steps: pill recognition (combined with pill classification) and pill retrieval. For pill recognition, we will focus on pill color, shape, size, and embossing to increase the accuracy of the model. Pill retrieval will then be performed using the identified features to retrieve the pill name from a set database.

As the first step in our pipeline, pill recognition, we plan to use YOLOv5 for initial object detection on input images. This enables the model to localize the pill and any embossing features on it. The studies that we researched in our literature review used YOLOv5 for their model implementation as well. The authors of those papers chose YOLOv5 for its real-time detection capability, high accuracy, and robustness to variations. We selected YOLOv5 for primarily the same reasons, and additionally for its ease of deployment, in order to better fit our time constraints due to the nature of this short-term project. We will be able to fine-tune efficiently to better fit our pill image dataset.

Following phase one of pill recognition, the identified pill regions and embossed characters will be passed through a ResNet-32 model for feature extraction, which is similar to the methodology of the studies we researched. The authors of those studies chose ResNet32 as it is able to dig deeper and identify distinguishing features such as shape and color of the pill to build more context for the model. This aligns with what we hope to accomplish with this model, and leverages what we have learned within this class so far.

For the pill retrieval step, the imprinted or embossed regions of the pill will be processed using a character language model for proper character recognition. We plan to experiment with both DeepTextSpotter and ASTER, which were used in the research within our literature review. Based on comparative results, we will select the model that handles text on curved surfaces or rotated images the best, and implement that model for our project. By choosing the best model for this method, we are compensating for the gap we identified in previous research where the models are not able to handle pictures of pills that are not oriented or pictured perfectly.

After using all three models to examine our images, the final characters will be put through an RNN to improve accuracy and refine the imprint output to match realistic scenarios. The outputs from the ResNet32 model, the character language model, and the final RNN model (if proven useful) will be concatenated to form a dataset that links the pill image features with the imprint text sequence. This concatenated dataset will be passed through an embedding space for similarity based matching, which will output the predicted pill ID based on which pill name fits the features and text the best. This method finds the top k matches, where k is a number we specify, of pill ID based on the input characteristics, with the top candidate returned to the user for final pill identification.

V. Experiments

To evaluate the performance of our models, we will run a series of experiments aimed at measuring both pill classification and imprint recognition accuracy. Evaluation metrics will include accuracy, precision, recall, and F1-score for each stage of the pipeline, as well as Intersection over Union for detection accuracy and character-level imprint recognition accuracy for text quality assessment. A confusion matrix will help visualize which pill classes are most frequently misclassified, in addition to providing insight regarding the general performance of the model.

We will also conduct an ablation study to isolate the contribution of each component. This will include comparing DeepTextSpotter and ASTER for text recognition and testing the pipeline with and without the RNN refinement layer. These comparisons will clarify which configurations provide the highest overall performance and robustness.

Finally, we plan to perform hyperparameter tuning using a combination of grid and random search to optimize learning rate, batch size, and optimizer selection. We may also explore different loss functions such as cross-entropy and CTC loss, to determine which best captures imprint recognition errors. Through this structured evaluation process, we aim to identify the model configuration that achieves the best balance of speed, accuracy, and generalizability across real-world pill identification scenarios.

References

[1] J. Heo, Y. Kang, S. Lee, D. H. Jeong, and K. M. Kim, "An Accurate Deep Learning-Based System for Automatic Pill Identification: Model Development and Validation," *Journal of Medical Internet Research*, vol. 25, p. e41043, 2023. [Online]. Available: https://doi.org/10.2196/41043

[2] N. Kavitha, A. Stefi, P. M. Varsha, S. Sumana, and V. V. Simha, "Real-time pill identification and classification using deep learning framework for medicine inspection system," Springer, 2025. [Online]. Available: https://doi.org/10.1007/s44291-025-00122-6

[3] T. Nguyen, T. Pham, N. Vu, and Q. Tran, "High accurate and explainable multi-pill detection framework with graph neural network-assisted multimodal data fusion," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2303.09782

[4] TruMedicines, "Pharmaceutical tablets dataset [Data set]," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/trumedicines/pharmaceutical-tablets-dataset