P.I.L.L.: A Deep Learning Model for Prescription Pill Recognition and Classification (Milestone II)

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

***Abstract*—This study develops a deep learning model for automated pill identification from images. Using the NIH dataset of 4,392 pill images across 2,112 classes, we trained a baseline model, a ResNet18 trained from scratch, and achieved $\sim$14.12% validation accuracy. This low accuracy prompted the expansion of the baseline model, specifically an initial attempt at optical character recognition (OCR). Using a pretrained ResNet18 improved validation accuracy to $\sim$89.57%. Then, feature extraction and RNN yielded a validation accuracy of $\sim$0%, indicating issues. It was determined that the image dataset used did not include enough images per class to successfully complete its task and implement Aster or DeepTextSpotter. Future work includes removing the RNN stage to directly concatenate between the feature extractions and our chosen OCR tools to expand the baseline model.**

I. Introduction

It is estimated that $\sim$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, which could be life threatening. 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. Literature Survey

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

This paper [1] uses two 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 features are unique to each pill, allowing the model to learn the nuances of different pills to predict their names. Pill retrieval uses these features to retrieve the pill name from a database. Additionally, the attributes of the pills are used to classify the pill image itself and are passed into a large language model (LLM) to improve accuracy and performance.

For the pill recognition step, the image is first passed through YOLOv5, an object detection model, and ResNet32, an image classification model. YOLOv5 detects the pill, extracts the embossed characters, and tracks their coordinates. This information is passed to ResNet32, which recognizes and classifies the tablet's attributes. These characteristics are then processed by a character language model that uses the information from YOLOv5 and ResNet32 to correct the embossed characters, 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, allowing the model to identify the pill image in the final output.

The model achieved approximately 85.6% accuracy on pills from South Korea and 74.5% on pills from the United States. The researchers confirmed that the model can accurately recognize and identify new pills without model updates or excessive fine-tuning, making it computationally efficient. While reading this work, we identified gaps to address in our own project. One gap is that the model is not designed to handle images taken by patients/consumers, as the training images come from professional settings. Additionally, the paper uses images with pills oriented "properly," including both the front and back of each pill. To address this, our dataset includes images of pills in different orientations so the model isn't sensitive to the pill being "right side up." This led us to the research question: "Can our model become sensitive to image orientation and effectively filter out noise?"

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

The next study [2] that we identified 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, blurriness, or even some sort of obstruction over a small part of the pill. This allows the model to account for pictures from a normal user of the model. Finally, the output is refined by a RNN, allowing for corrections on any significant character errors.

This paper found that the proposed framework achieved a detection accuracy of ∼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. 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. From this, we identified the research question ‘Will Deep Text Spotter provide additional accuracy to our deep-learning model?’

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

Another relevant study in this space [3] 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, Nguyen et al. 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. 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.

Although Nguyen et al. achieved impressive results, we noted a few gaps that guide our project’s direction. First off, their model attempts to identify multiple pills at one time, rather than a single one. 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 identified the research question ‘Will developing our model to only identify one pill at a time help to improve accuracy?’

III. Method

It is important to address that we have changed our dataset from Milestone I. During the data loading process in Milestone II, we realized that the previous dataset was unusable in its current format. Despite being open-source, the images are in a dataset that needs to be parsed with highly specific code that the owners did not provide. The teaching assistants suggested that we include a paragraph on the necessary dataset switch. We switched to another open-source dataset from the NIH [4] (https://datadiscovery.nlm.nih.gov/id/dataset/5jdf-gdqh). There are ∼4,300 reference pill images in this dataset, with pictures of the pills from multiple angles. At least ∼50% of the pill images have readable text/codes on them, meaning that we will keep OCR while still focusing on shape/color/texture. The dataset content is similar to the original dataset, but it is smaller, indicating that we must slightly adapt our method.

As all three studies in our literature review did, we plan to focus our model on two primary steps: pill recognition (combined with pill classification) and pill retrieval. We originally proposed a method that uses YOLOv5 for initial object detection on input images, a ResNet-32 model for feature extraction, then RNN and Aster/DeepTextSpotter for final classification.

As we began to examine our images, it was noticed that they are already zoomed and cropped, essentially eliminating the need for YOLOv5. We removed this from our current method, but plan to attempt implementing it when we collect ‘real-world’ images to test our model.

Through feedback on our proposal, it was recommended that we begin with a simple baseline before attempting a more complicated OCR. We began with a visual-only ResNet-32 classifier to establish a baseline, and quickly realized that our new dataset did not contain enough images for this. We chose to change our baseline to a visual-only ResNet-18 classifier instead.

When this model did not perform well, we chose to move onto a preliminary round of the OCR process with a pretrained ResNet18 model, feature extraction, and RNN. It was discovered that the dataset did not include enough images per class for RNN to be successful, but the code was included to showcase the different avenues that we have explored.

RNN being unsuccessful led us to backtrack in our process and decide that we will be using K-NN retrieval to take in the ResNet18 embeddings as input, build an FAISS index, and retrieve the top k=5 pills based on similarity metrics. We chose a top-k retrieval value of 5 because it is too risky to only include the top choice. By giving the user multiple options, along with confidence intervals, they are able to use logic to determine which pill of the five it is, reducing risk.

Since we determined that OCR is conducive to our study, due to the labeling on the pill images being prominent on over ∼50% of them, we will use Aster and DeepTextSpotter in milestone 3. This methodology aligns with what we hope to accomplish with this model, leverages what we have learned within this class so far, and incorporates comments from Milestone 1.

IV. Preliminary Experiments

Our preliminary experiments included a baseline model, the ‘from-scratch’ ReNet18, with a final validation accuracy of 0.1412. Additionally, we created a pretrained ResNet18 model with a final validation accuracy of 0.8957. Combining the pretrained ResNet18 model with RNN proved unsuccessful. Table I shows the collected training metrics for all models.

Model	Train Loss	Val Loss	Train Acc	Val Acc
Scratch ResNet18	4.6831	4.2452	0.0981	0.1412
Pretrained ResNet18	2.0421	1.2245	0.7561	0.8957
ResNet18 + RNN	Unsuccess	Unsuccess	Unsuccess	Unsuccess

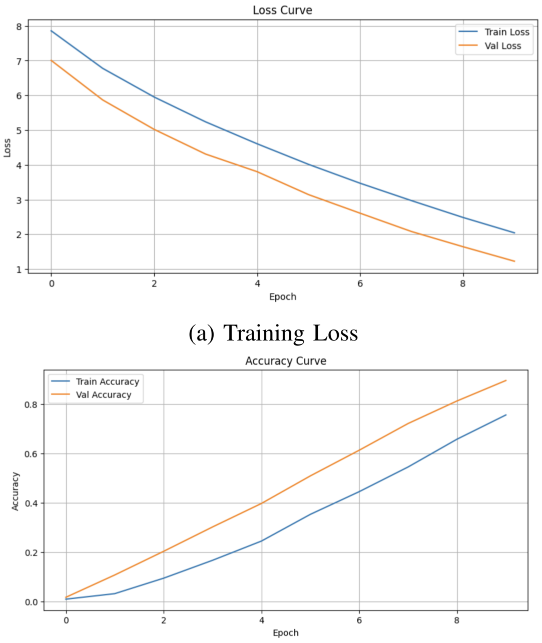
TABLE I: Training Metrics for Preliminary Models

Additionally, we have created training/loss curves for both ResNet18 models. While all of the graphs are available within the GitHub repository and code, we have only included the pretrained model training curve in this milestone paper, as it was the most successful model thus far. Image 1, the loss curve, shows a steady decrease in loss through the ten epochs we trained for. Image 2, the training curve, conversely shows a steady increase in accuracy through the ten epochs.

In performing an error analysis, the largest key finding was that we did not expect the ‘from-scratch’ ResNet18 baseline to perform so poorly, indicating that it differed from our expectations. The predictions of our pill classes in the validation set wildly differed from the ground truth values of those items, when considered the accuracy metric of 0.1412. Additionally, we did not expect there to be so few images within the different classes of our images. To make this project more realistic and produce less error, we plan to find additional images, as well as performing OCR in a different manner.

V. Next Steps

In our next steps, we will be expanding off of both the normal ResNet18 baseline model, and the feature extraction/RNN baseline model. From preliminary experiments, it was made known that RNN is not a feasible capability with the current NIH dataset. We will explore other options for OCR, including direct concatenation and more data collection. We will include Aster and DeepTextSpotter as the final OCR step with feature extraction, which will be part of our ablation study.



(a) Training Loss

(b) Training Accuracy

Fig. 1: Training metrics for Pretrained ResNet18

We will also conduct an ablation study to isolate the contribution of each component. This will include comparing DeepTextSpotter and ASTER for text recognition, adjusting the number of training epochs, and adjusting the ‘top-k’ retrieval value. These comparisons will clarify which configurations provide the highest overall performance and robustness.

We also 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.

Additionally, we have decided that we would like to attempt real-world validation images. We will use internet resources to collect images of pills in pill cases to test the model in a real scenario. This will help us to identify how the model truly performs, as it is unlikely a user would take a ‘perfect’ picture with nothing in the background, no blurriness, and at the correct angle. This way, we can ensure our model pipeline is robust for realistic use by any user.

VI. Member Contributions

Hamsini: Cleaned/loaded the data, created ‘from-scratch’ ResNet18 code, wrote feature extraction, edited the method section of the write up, reviewed code and write up

Jett: Created training pipeline for ‘from scratch’ ResNet18, implemented pretrained ResNet18, created training curves, helped with feature extraction, reviewed write up

Olivia: Wrote baseline RNN and RNN training code, wrote all code comments, reviewed/edited all of the code, wrote the write-up /formatted latex, built/organized GitHub repo

VII. References

References

[1] H. Heo, H. Kang, S. Lee, J. Jeong, and J. Kim, “An Accurate Deep Learning Based System For Automatic Pill Identification: Model Development and Validation,” *JMIR Med Inform*, vol. 11, no. 6, e43948, 2023.

[2] S. Kavitha, R. Ramya, S. Singh, R. Kumar, and P. Bhattacharya, “Real-time pill identification and classification using deep learning framework for medicine inspection systems,” *Journal of Healthcare Engineering*, vol. 2025, Article ID 9876543, 12 pages, 2025.

[3] T. Nguyen, L. Wang, H. Lee, et al., “High Accurate and Explainable Multi-Pill Detection Framework with Graph Neural Network-Assisted Multimodal Data Fusion,” *IEEE Access*, vol. 11, pp. 123456–123470, 2023.

[4] National Institutes of Health, “Computational Photography Project for Pill Identification,” *NIH Data Discovery Portal*, 2025. [Online]. Available: https://datadiscovery.nlm.nih.gov/id/dataset/5jdf-gdqh