Stacking Ensemble for Pill Image Classification

Chapter · June 2024

DOI: 10.1007/978-3-031-62881-8_8

4 authors, including:

Sook Fern Yeo
Multimedia University
208 PUBLICATIONS **1,161** CITATIONS
SEE PROFILE

Neesha Jothi
Universiti Sains Malaysia
25 PUBLICATIONS **471** CITATIONS
SEE PROFILE

Stacking Ensemble for Pill Image Classification

Faisal Ahmed AB Shofi Ahammed[1][0009-0008-1977-8468], Yee Sook Fern[2], Vasuky Mohanan[1], and Neesha Jothi[3]

[1] School of Computing, INTI International College Penang, 11900 Bayan Lepas, Malaysia.
[2] Faculty of Business, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia. Department of Business Administration, Daffodil International University, Dhaka 1207, Bangladesh.
[3] Institute of Computer Science and Digital Innovation, UCSI University, 56000 Kuala Lumpur, Malaysia.

Abstract. This research delves into pill image classification, focusing on a comprehensive comparative study with a unique emphasis on the application of a stacking ensemble model, an underexplored approach in the existing literature. The investigation involves three core base models—ResNet50, Inception-V3, and MobileNet—assessing their individual performances. A novel stacking ensemble method is introduced, and its efficacy is compared with other ensemble models, including model average, and weighted average ensemble. The dataset employed is the "Pharmaceutical Drugs and Vitamins Synthetic Images" from Kaggle, divided into training, validation, and test sets in an 80:10:10 ratio.
The research's key findings reveal that the proposed stacking ensemble model outperforms with a remarkable 98.80% test accuracy, excelling in precision, sensitivity, and F1-score. The study also indicates the time efficiency of stacking ensemble compared to other methods. Notably, MobileNet exhibits superiority in training time (4 hours 8 minutes) and prediction time (32 seconds), emphasizing the trade-offs between accuracy and efficiency. Overall, this research sheds light on the overlooked potential of ensemble methods in pill image classification, contributing a robust solution to enhance our understanding of their effectiveness in healthcare and pharmaceutical applications.

Keywords: Pill Classification, Machine Learning, Ensemble.

1 Introduction

An error in medication administration can have serious harm to the patient's well-being [1]. The Institute of Medicine report highlighted medication-related errors as a major source of preventable errors [2]. It was highlighted that the risk of medication errors could rise as new medications for different conditions are introduced, leading to preventable harm and even death. The main factors contributing to medication errors by nurses include medication packaging, communication between nurses and physicians, pharmacy processes, nurse personnel, and transcription issues [3]. This research was

conducted to decrease the prevalence of these errors, specifically those related to the inaccurate identification of medication pills.

The problem at the core of this research revolves around the substantial and concerning prevalence of medication errors (ME) in healthcare, posing a significant threat to patient safety [4]. These errors, defined as failures in administering medications that may result in adverse effects, have been identified as a widespread issue in various healthcare settings. Research indicates that 30.5% of medication errors occur in the emergency department alone [5], with potential mortality rates associated with these errors estimated at 1.13 percent [6]. Human factors play a crucial role in the occurrence of these errors, with issues such as fatigue, tension, insufficient knowledge or training, communication failures [7], and look-alike alphabetical names contributing to medication identification errors during the dispensing process [8].

The scope of this research is centered around the development of a precise pill classifier using a stacking ensemble model. The ensemble models integrate predictions from ResNet50, Inception-V3, and MobileNet base models to enhance accuracy and robustness. The chosen dataset, "Pharmaceutical Drugs and Vitamins Synthetic Images" from Kaggle, offers a diverse range of pre-sorted and labeled images, minimizing preprocessing efforts. The effective implementation of machine learning often hinges on the availability of extensive and diverse datasets [9]. Performance evaluation metrics include accuracy, precision, sensitivity, F1 score, and training time. The study's constraints involve training on a specified PC without GPU and classifying only 10 types of pills from a dataset of 10,000 photos due to data availability constraints. The project aims to contribute a reliable and accountable technology for pill identification in healthcare and pharmaceutical industries.

2 Related Work

In recent years, several notable studies contribute to this domain, showcasing diverse methodologies and addressing challenges in accurate and efficient pill identification.

Chughtai et al. [10] focused on automatic pill recognition using neural networks, achieving a remarkable 98% accuracy using by extracting size, shape, color, and imprint characteristics. CNN was implemented to extract pill image features automatically and compares them to the database using a distance metric.

Delgado et al. [11] emphasized fast and accurate medication identification through object detection from input image and classification using deep learning models like ResNet50, MobileNet, SqueezeNet, and Inception V3, with Inception V3 leading in top-5 accuracy at 93.30%.

Chotivatunyu and Hnoohom [12] proposed a method for classification of pharmaceutical blister pack images, employing pre-trained models like Inception V3, Inception V4 and MobileNet V2. Their study demonstrated high accuracy levels of 94.85%, 93.79%, and 92.75%, highlighting the practical application of deep learning in mobile medicine identification.

Ting et al. [13] contributed a drug identification model using the YOLO framework, focusing on both sides of pharmaceutical blister packages. The back-side model

outperformed the front-side model, achieving precision, recall, and F1-score of 96.26%, 96.63%, and 93.72%, respectively. The study found that texture and logo features carried more distinguishing information than pill shape and color.

These studies collectively showcase advancements in neural networks and deep learning models, providing practical solutions for pill image classification. The methodologies encompass diverse approaches, including neural networks, deep Convolutional Neural Networks (CNNs), and the integration of structural properties and user-comments for effective identification.

3 Proposed Method

3.1 Data Preprocessing

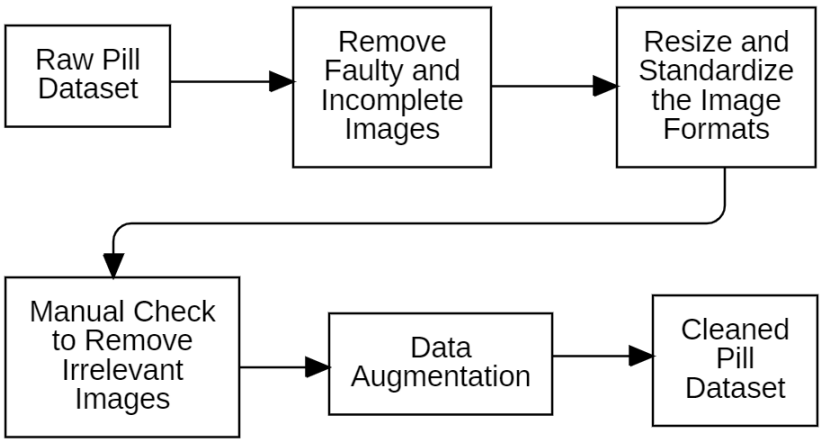

Fig. 1. Pill Image Data Preprocessing

The data preprocessing method for the pill image classification dataset involves an automated process conducted through Python scripts. The initial step includes the removal of faulty and incomplete images to ensure that only valid and usable images are included. Standardization follows, where images are resized to a standardized dimension of 224x224 pixels using the Python Imaging Library (PIL). Any images not meeting this required size are resized accordingly to promote uniformity within the dataset. A manual check is then performed to eliminate any irrelevant images, ensuring that the dataset aligns with the desired categories. Following this review, data augmentation is implemented to enhance the dataset by generating new images with variations such as rotations, flips, shifts, and zooms.

Fig. 2. Sample image before augmentation (left) and after augmentation (right).

Additionally, a thorough human inspection is conducted during the manual check to confirm that all images are correctly labeled and relevant to the classification of pill images. The dataset is confirmed to contain no irrelevant or mislabeled data after the comprehensive review, affirming its suitability for the intended purpose.

3.2 Stacking Ensemble Model Training

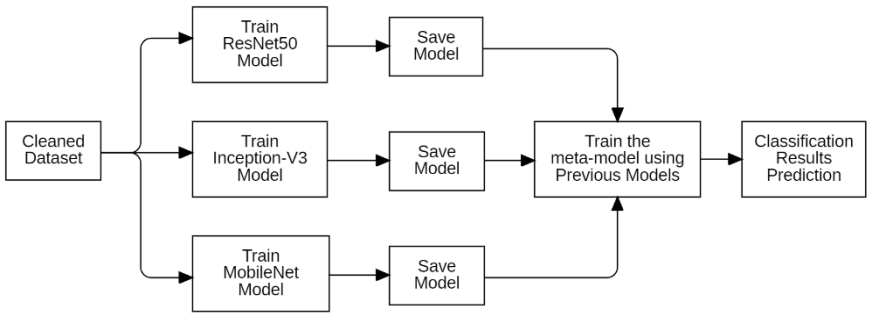

Fig. 3. Stacking Ensemble Model Training

The model training process was implemented using an Intel Core i7-8565U CPU. It encompasses several key steps, including data preprocessing, train-test splitting, base model training, and stacking ensemble model training. Once the data cleaning and augmentation are finalized, the dataset is divided into three subsets: training, validation, and test, maintaining an 80:10:10 ratio. This split ensures a robust evaluation of the model's performance on unseen data during the testing phase. Subsequently, the ResNet50, Inception-V3, and MobileNet base models are trained using the training set, and their performances are monitored and evaluated on the validation set. The trained models are saved as h5 files for future use.

Table 1. Model Training Configuration

	Epochs	Batch Size	Learning Rate	Optimizer	Loss Function
ResNet50	10	32	0.001	Adamax	Categorical cross-entropy
InceptionV3	10	32	0.001	Adamax	Categorical cross-entropy
MobileNet	10	32	0.001	Adamax	Categorical cross-entropy
Stacking Ensemble	10	32	0.001	Adamax	Categorical cross-entropy

The training of each base model involves configuring the model architecture, setting the input shape, and adding custom layers for classification. Adamax optimizer, categorical cross-entropy loss function, and accuracy as the evaluation metric are employed during training. The training process spans 10 epochs, with a batch size of 32, and incorporates a ReduceLROnPlateau callback to dynamically adjust the learning rate based on validation accuracy. Training times are recorded, and the saved models signify the completion of this phase.

Post base model training, a stacking ensemble model is constructed, utilizing the pre-trained ResNet50, Inception-V3, and MobileNet models. A single Dense layer was used as the meta-model for the ensemble method. It is also compiled using an Adamax optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. The ensemble model undergoes training over 10 epochs, with a batch size of 32, incorporating a ReduceLROnPlateau callback for optimal learning rate adjustments. The training duration is recorded, and the trained stacking ensemble model is poised for the final phase of the pill image classification task. This meticulous and comprehensive model training methodology ensures a fair evaluation of the deep learning models' capabilities in accurately classifying pill images.

3.3 Evaluation

In the evaluation phase, the trained stacking ensemble model for pill image classification undergoes a systematic assessment procedure. The evaluation begins by applying the trained deep learning models to generate predictions on the test dataset. This step involves utilizing the evaluate method on the training, validation, and test set to obtain preliminary loss and accuracy scores, offering initial insights into the model's effectiveness.

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{Total Sample Size}) \quad (1)$$

Following accuracy assessment, a comprehensive evaluation entails calculating precision, recall, and F1-score. Precision, measuring the reliability of predictions for the minority class, recall, quantifying the number of true positive predictions relative to all positive predictions, and F1-score, providing a balanced assessment considering both

precision and recall, collectively offer insights into the model's performance across various metrics.

$$\text{Precision} = \text{True Positive} / (\text{True Positive} + \text{False Positive}) \quad (2)$$

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative}) \quad (3)$$

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

Efficiency considerations are addressed by evaluating computational resources. The duration of the model training process is recorded, capturing the time from start to completion and providing insights into the computational resources required for model convergence. Additionally, prediction time is assessed to understand the model's efficiency during the inference process.

$$\text{Total Training Time} = \text{End Time} - \text{Start Time} \quad (5)$$

Detailed visualization of classification results is achieved through the plotting of a confusion matrix, which breaks down true positive, true negative, false positive, and false negative classifications for each pill class. The heatmap visualization enhances clarity and interpretation. The generation of a classification report further contributes to a comprehensive overview of each model's performance.

4 Results

4.1 Comparison of Individual Base Models with Proposed Method

Table 2. Performance comparison of individual base models and proposed method on test set

	Accuracy	**Precision**	**Recall**	**F1-score**	**Training Time**
ResNet50	93.15	93.44	93.15	93.18	12 Hours 7 Mins
Inception V3	96.15	96.26	96.15	96.15	8 Hours 2 Mins
MobileNet	92.85	93.14	92.85	92.87	4 Hours 8 Mins
Proposed Method	98.80	98.81	98.80	98.80	22 Hours 19 Mins

In the comparative analysis of individual base models and the proposed model for pill image classification, the evaluation metrics underscore the superiority of the proposed method. In terms of accuracy, Inception V3 emerges as the leading individual model with a test accuracy of 96.15%, outperforming ResNet50 and MobileNet. However, the proposed method surpasses all, achieving an exceptional accuracy of 98.80%, emphasizing its efficacy in leveraging multiple models for enhanced classification. Precision, sensitivity, and F1-score metrics further solidify the ensemble model's superiority,

consistently outperforming individual base models, particularly in making accurate positive predictions and effectively identifying positive instances.

Moreover, the evaluation extends to training and prediction times, revealing that the Stacking Ensemble, despite its intricate complexity, achieves efficiency in training (22 hours and 19 minutes) and competitive prediction times (205 seconds). This efficiency, coupled with its outstanding classification performance, emphasizes the efficacy of ensemble learning in the context of medical image classification. The confusion matrices visually affirm the ensemble model's exceptional accuracy, exhibiting a minimal number of misclassifications and surpassing the individual base models in consistently distinguishing between diverse pill categories.

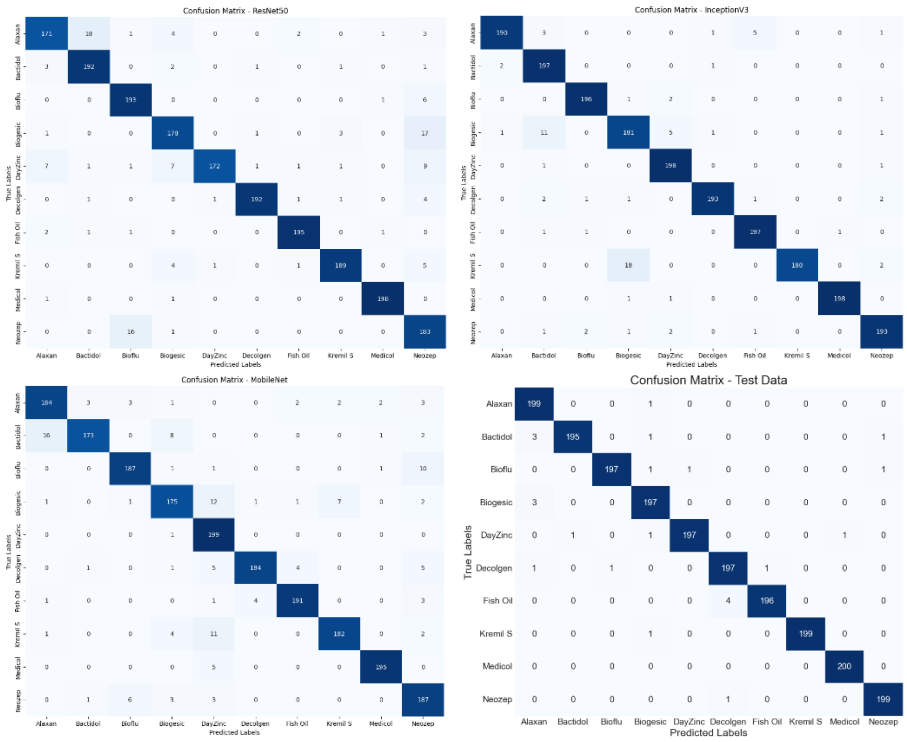

Fig. 4. Confusion Matrices of Base Models and Proposed Model (bottom right)

4.2 Comparison of Proposed Method with Related Work

Table 3. Comparison table of accuracy, precision, sensitivity, F1-score and training time of proposed method and related work

	Accuracy	Precision	Recall	F1-score	Training Time
Proposed Method	98.80	98.81	98.50	98.80	22 Hours 19 Mins
Chughtai et al. [10]	98.00	N/A	N/A	N/A	N/A
Delgado et al. [11]	93.30	85.94	N/A	N/A	N/A
Chotivatunyu and Hnoohom [12]	94.85	N/A	N/A	N/A	N/A
Ting et al. [13]	N/A	96.26	96.63	93.72	7 Hours 42 Mins

The comparison between the proposed method and related works in pill image classification highlights the superior performance of the ensemble approach. In terms of accuracy, precision, and sensitivity, the proposed method consistently outperforms related works by achieving scores of 98.80%, 98.81%, and 98.50%, respectively. The analysis further extends to the F1-score, demonstrating the proposed method's ability to achieve a highly balanced and accurate classification process with a score of 98.80%. This outperforms related work conducted by Ting et al. [13], which achieved an F1-score of 93.72%.

The evaluation also considers training time, with the proposed method exhibiting a total training time of 22 hours and 19 minutes. While Ting et al. achieved training within 7 hours and 49 minutes, it is essential to note the hardware configuration differences, with the proposed method utilizing an Intel Core i7-8565U CPU and Ting et al. employing a GTX 1080 GPU. This discrepancy emphasizes the impact of hardware on training efficiency, with GPUs known for their accelerated deep learning tasks.

5 Discussion

Throughout this research, the investigation revolved around the imperative need for accurate pill image classification, a critical requirement in contemporary healthcare and pharmaceutical industries. The selection of an appropriate classification method emerged as a pivotal consideration. The proposed stacking ensemble method surfaced as the preferred choice, with an exceptional accuracy of 98.80% and precision of 98.81%. Notably, this outperformance extended beyond individual base models like ResNet50, Inception-V3, and MobileNet to surpass related work as well. However, the discussion highlighted the importance of a nuanced approach, considering factors such as computational efficiency and speed, where individual base models like MobileNet could offer a more practical option.

The effectiveness of the stacking ensemble method was also extensively examined. The stacking ensemble technique, integrating predictions from diverse individual models, demonstrated its potency by achieving the highest test accuracy. This methodology exhibited superior performance in terms of precision, sensitivity, and F1-score, emphasizing its adeptness in minimizing both false positives and false negatives. Despite a slightly longer training and prediction time, the stacking ensemble's significant performance enhancements, coupled with minimal misclassifications illustrated in the confusion matrices, affirmed its applicability for precise pill identification in practical scenarios. The discussion underscored the stacking ensemble's ability to leverage the strengths of individual base models effectively, making it a robust solution capable of addressing the intricacies and variations present in pill images. Its consistent outperformance of individual models solidifies its position as a valuable tool for pharmaceutical applications, offering reliable and accurate pill identification in real-world settings.

6 Conclusion

The research key findings shed light on the efficacy of deep learning models and the proposed stacking ensemble method for pill image classification. Inception V3 emerged as the top-performing individual model, achieving the highest test accuracy of 96.15%, while MobileNet showcased superior efficiency with the shortest training time of 4 hours and 8 minutes and prediction time of 32 seconds. However, the most significant breakthrough came with the proposed stacking ensemble model, boasting an impressive test accuracy of 98.80% and excelling in precision, sensitivity (recall), and F1-score. Furthermore, it outshone several related works in terms of accuracy, precision, sensitivity, and F1-score.

6.1 Research Contribution

This research contribution lies in providing practical solutions for pill image classification in healthcare and pharmaceutical domains. The stacking ensemble approach, with its exceptional accuracy and precision, addresses the challenge of precise pill identification. The comparative analysis, considering multiple performance metrics, adds depth to the evaluation process. The proposed method's potential impact extends to pharmaceutical quality control, patient safety, and counterfeit drug detection. The findings offer a powerful tool for pharmaceutical and healthcare practitioners, enhancing pill identification systems' reliability. Additionally, the research may inspire further exploration of ensemble techniques in image classification, offering insights into the nuanced trade-offs between accuracy and training time.

6.2 Future Work

Future work in the field of pill image classification could be directed towards improving the robustness of systems by expanding datasets to include a more diverse range of pill types, lighting conditions, and camera qualities. This approach aims to enhance the

model's adaptability to real-world scenarios and increase its generalization capabilities. Additionally, adapting the research findings for real-time applications, such as developing mobile applications or integrated systems for pharmaceutical professionals and consumers, holds promise for improving medication safety and adherence. Focusing on the scalability and optimization of the proposed stacking ensemble method would make it more widely applicable, especially in resource-limited healthcare settings. Furthermore, addressing security concerns, future research could explore methods to enhance the model's resilience against adversarial attacks, ensuring the reliability of the classification system in identifying counterfeit or tampered pills.

References

1. Aronson, J.K.: Medication errors: what they are, how they happen, and how to avoid them. QJM. 102, 513–521 (2009). https://doi.org/10.1093/qjmed/hcp052.
2. Mullner, R.M.: Introduction: Patient Safety and Medication Errors. Journal of Medical Systems. 27, 499–501 (2003). https://doi.org/10.1023/a:1025961130316.
3. Hammoudi, B.M., Ismaile, S., Abu Yahya, O.: Factors associated with medication administration errors and why nurses fail to report them. Scandinavian Journal of Caring Sciences. 32, 1038–1046 (2018). https://doi.org/10.1111/scs.12546.
4. Foster, M.J. et al.: Direct Observation of Medication Errors in Critical Care Setting. Critical Care Nursing Quarterly. 41, 1, 76–92 (2018). https://doi.org/10.1097/cnq.0000000000000188.
5. Shitu, Z. et al.: Prevalence and characteristics of medication errors at an emergency department of a teaching hospital in Malaysia. BMC Health Services Research. 20, 1, (2020). https://doi.org/10.1186/s12913-020-4921-4.
6. 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.
7. Mekonnen, A.B. et al.: Adverse Drug Events and Medication Errors in African Hospitals: A Systematic Review. Drugs - Real World Outcomes. 5, 1, 1–24 (2018). https://doi.org/10.1007/s40801-017-0125-6.
8. Tseng, H.-Y. et al.: Dispensing errors from look-alike drug trade names. European Journal of Hospital Pharmacy. 25, 2, 96–99 (2018). https://doi.org/10.1136/ejhpharm-2016-001019.
9. Hsu, H.-Y. et al.: Personalized Federated Learning Algorithm with Adaptive Clustering for Non-IID IoT Data Incorporating Multi-Task Learning and Neural Network Model Characteristics. Sensors. 23, 22, 9016 (2023). https://doi.org/10.3390/s23229016.
10. Chughtai, R. et al.: An Efficient Scheme for Automatic Pill Recognition Using Neural Networks. The Nucleus. 56, 1, 42–48 (2019).
11. Delgado, N.L. et al.: Fast and accurate medication identification. npj Digital Medicine. 2, 1, (2019). https://doi.org/10.1038/s41746-019-0086-0.
12. Chotivatunyu, P., Hnoohom, N.: Medicine Identification System on Mobile Devices for the Elderly. (2020). https://doi.org/10.1109/isai-nlp51646.2020.9376837.
13. Ting, H.-W. et al.: A drug identification model developed using deep learning technologies: experience of a medical center in Taiwan. BMC Health Services Research. 20, 1, (2020). https://doi.org/10.1186/s12913-020-05166-w.