

Classification of Crystallization Images of Pharmaceutical Raw Materials Using Convolutional Neural Network Algorithm

Julia Mega Reski¹, Anton Yudhana²

¹Master Program in Electrical Engineering, Universitas Ahmad Dahlan, Indonesia

²Department of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:

Received Jun 29, 2025

Revised Oct 21, 2025

Accepted Nov 12, 2025

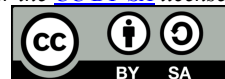
Keywords:

CNN,
crystallization,
deep learning,
pharmaceutical raw materials,
image classification

ABSTRACT

The rapid advancement of artificial intelligence (AI) has opened new opportunities for automation in the pharmaceutical industry, particularly in the classification of raw drug materials. Manual classification methods are time-consuming and prone to human error, highlighting the need for reliable automated solutions. This study applied a deep learning approach for classifying crystallization images of pharmaceutical raw materials using a Convolutional Neural Network (CNN). A dataset of 300 crystallization images of Nicotinamide and Ferulic Acid was obtained through hot-stage microscopy, preprocessed with normalization, resizing, and augmentation, and divided into training, validation, and testing subsets. The CNN model was trained for 10 epochs and evaluated using a confusion matrix and standard performance metrics (accuracy, precision, recall, and F1-score). The model achieved perfect recall for Ferulic Acid and 90% recall with 100% precision for Nicotinamide, resulting in an overall accuracy of 95%. While these results are promising, the relatively small dataset may limit generalization, and further validation with larger or external datasets is required. The findings indicate that CNN-based methods hold strong potential for automating crystallization classification, improving pharmaceutical quality control, and reducing reliance on manual assessment, in line with recent advances in medical and pharmaceutical image analysis.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Anton Yudhana

Department of Electrical Engineering

Universitas Ahmad Dahlan

Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Bantul, Daerah Istimewa Yogyakarta 55191

Email: eyudhana@ee.uad.ac.id

1. INTRODUCTION

Manual classification of pharmaceutical raw materials is time-consuming and prone to human error, making it a significant challenge in drug production [1]. A thorough understanding of physicochemical properties, such as melting point and crystallization behavior, is critical for ensuring product quality [2]. The melting point serves as a key indicator of purity, while recrystallization after melting often produces new crystal structures with distinct characteristics shape, size, and polymorphic form that influence stability, solubility, and bioavailability. For example, carbamazepine recrystallization has been shown to generate new polymorphs with improved dissolution rates [3], while microfluidic studies have demonstrated that cooling rates affect nucleation and polymorphism in APIs such as irbesartan and aripiprazole [4]. These findings underscore the importance of strict standards in pharmaceutical production to maintain quality and consistency [5].

To analyze such phenomena, this study employed hot-stage microscopy (HSM), which integrates thermal analysis with real-time microscopic observation, enabling detection of

polymorphism, phase transitions, and morphological changes not captured by DSC or TGA [6]. The crystallization images produced were then classified using Convolutional Neural Networks (CNNs), a widely adopted deep learning method known for effectively extracting complex visual features [7], [8], [9]. CNN-based systems such as CrystalNet and MARCO have achieved high accuracy in crystal recognition, but their application in pharmaceutical raw material characterization remains limited [10]. Moreover, the dataset in this study, consisting of 300 crystallization images of Nicotinamide and Ferulic Acid is relatively small, raising concerns about generalization and robustness.

Previous works have also compared CNNs with traditional classifiers, reporting CNN's superior performance in tasks such as Pap smear image classification ($\approx 93\%$ testing accuracy versus 85% for SVM and KNN) [11], and digit recognition on MNIST and COREL1000 datasets (0.98 versus 0.88 accuracy for SVM) [12]. However, many studies suffer from small datasets, lack of external validation, and insufficient methodological detail, limiting reproducibility and generalization. Addressing these gaps, this research applies a CNN-based approach for classifying pharmaceutical crystallization images while explicitly describing preprocessing, augmentation, and training configurations. The objective is to evaluate CNN feasibility in pharmaceutical crystallization classification and to contribute toward developing reliable automated systems, while also acknowledging the limitations of working in data-constrained scenarios.

2. RESEARCH METHOD

The general system design is illustrated in Figure 1, which shows the overall research framework. This study is divided into six parts: data collection, data preprocessing, data splitting, model training, model testing, and evaluation.

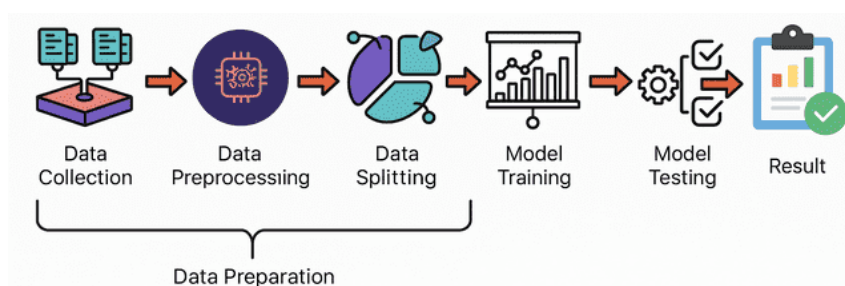


Figure 1. Research Framework

Figure 1 illustrates the process of classifying crystallization images of pharmaceutical materials using CNN, which begins with data collection in this case, crystallization images of drugs such as Nicotinamide and Ferulic Acid as the main dataset. Next, the data undergo preprocessing, which includes image normalization, noise removal, resizing, and data augmentation, in line with practices described in studies on deep learning-based crystal detection [13]. After that, the dataset is divided through data splitting.

At the model training stage, CNN is trained using the training data to recognize specific patterns in crystallization images, such as texture, shape, and crystal structure. A similar approach has been applied in the context of protein crystallization image classification, which compared several popular CNN architectures and emphasized the importance of efficient training strategies and feature representation[14]. After the model is trained, the next stage is model testing, where its performance is evaluated using previously unseen test data. This stage leads to the final results in the form of model performance metrics, consistent with the practical CNN pipeline methodology proposed in that study.

2.1 Acquisition of Original Crystallization Images with a Size of 350 x 350 Pixels

The data used in this study are crystallization images of the pharmaceutical materials Nicotinamide and Ferulic Acid with a size of 350 x 350 pixels, consisting of 300 image samples.

The crystallization images were obtained from the crystallization process of Nicotinamide and Ferulic Acid after melting, using a Hot-stage Microscope. The crystallization image data represent pure materials that were not contaminated, as evidenced by the melting points of the two pharmaceutical substances matching the dataset. Nicotinamide melts within the temperature range of 128–131°C, while Ferulic Acid melts at 168–170°C [15]. The steps for data collection are shown in Figure 2.

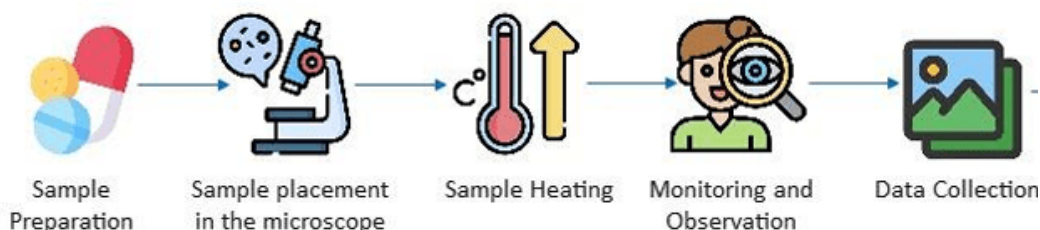


Figure 2. Data Collection Process

Figure 2 shows the workflow of crystallization image data collection for pharmaceutical materials. The first stage is sample preparation, in which Nicotinamide and Ferulic Acid are prepared in pure form. Next, the samples are placed on a hot-stage microscope (HSM), a technique that combines thermal analysis and microscopy to enable real-time observation of physical changes such as melting, recrystallization, polymorphism, and other solid-state morphologies [6].

A controlled heating process was applied to reach the temperature that triggers the physical transformation of the sample, which was then observed through a polarization microscope. Following this, during the gradual cooling process, crystallization patterns are confirmed through optical observation. The image data obtained from heating and crystallization are then collected as a dataset for CNN model training and testing. Such an approach has been explored in studies of cocrystal characterization, where HSM is used as a primary tool in thermal evaluation [16].

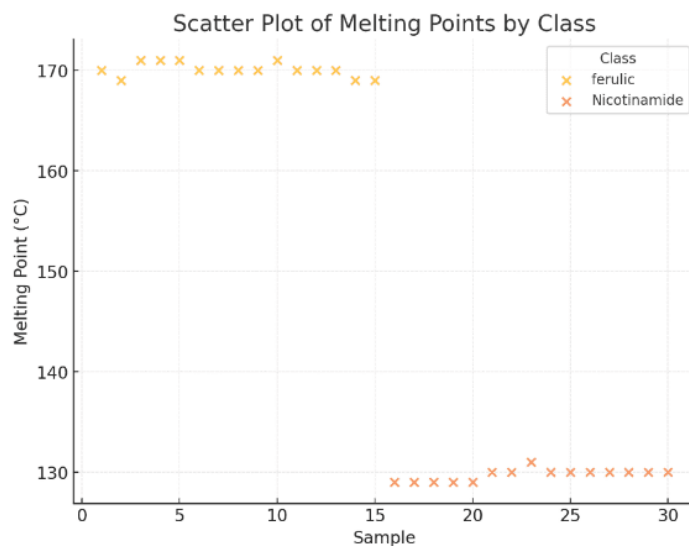


Figure 3. Melting Point Distribution of Ferulic Acid and Nicotinamide

Figure 3 shows the distinct difference in melting points. The Ferulic Acid class most frequently exhibits a melting point at 170°C, indicating that the majority of data in this class are concentrated around that value with a narrow and consistent distribution. In contrast, the Nicotinamide class has a much lower most frequent melting point, around 130°C, with a distribution that is also fairly consistent, although slightly narrower.

2.2 Data Preprocessing

Preprocessing is an essential step in digital image processing, aimed at improving visual quality, reducing noise, and preparing images to be optimal for analysis or deep learning model training (ensuring homogeneous image presentation) [17]. Common techniques in this stage include resizing (standardizing dimensions), format conversion, pixel normalization, and image restoration [16]. In particular, resizing plays a crucial role in aligning input sizes so that the model can extract features more consistently and accelerate the training process [18].

To improve robustness and reduce overfitting, we applied on-the-fly data augmentation exclusively to the training set using Keras' ImageDataGenerator. All input images were first normalized by rescaling pixel intensities to $[0,1][0,1][0,1]$ (rescale = $1/255$) and resized to the network input size ($128 \times 128 \times 3$). A horizontal flip (horizontal_flip = True) was enabled to simulate orientation variability commonly observed in crystallization micrographs while preserving the original crystal morphology. Augmented batches were generated with flow_from_dataframe, using the file paths and labels stored in a dataframe, with shuffling enabled to decorrelate successive mini-batches.

In preliminary experiments we considered additional geometric transforms (rotation, translation, shear, and zoom), but these were not retained in the final configuration to avoid distorting fine-grained crystal textures that are critical for class discrimination. This setup prevents leakage by applying augmentation only after the train/validation/test split, and only to the training pipeline.

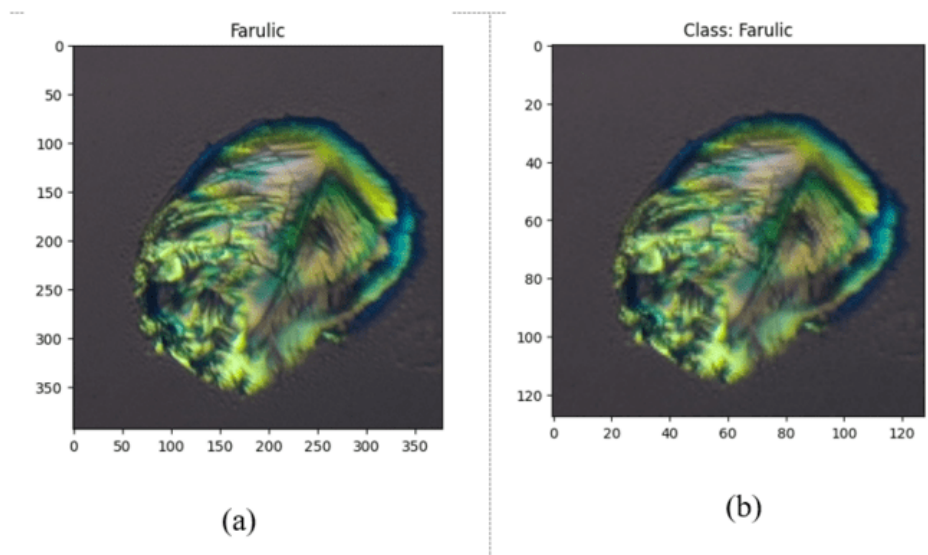


Figure 4. Images (a) Before and (b) After Preprocessing

2.3 Data Splitting

The data in this study were divided into three subsets: training data, used to guide the model in learning the main patterns from the dataset; validation data, which plays a role in hyperparameter tuning and detecting overfitting a condition where the model over-optimizes on the training data, resulting in decreased performance on new data. The causes and mitigation strategies for this issue have been systematically reviewed in the context of neural networks. In addition, the quality of data splitting (e.g., randomness and fair distribution) directly affects the validity of the final model evaluation, both for generalization and performance estimation, as explained by [19] through a diagnostic approach to data partitioning. Test data are only used after the entire training process is complete, to provide an objective picture of the model's ability to recognize previously unseen data. This is critical, especially for deep learning architectures that are prone to inflated false accuracy due to data leakage [20], [21]. In this study, the dataset was divided into three parts: 60% training data, 20% validation data, and 20% testing data, with a total of 300 images. The current

evaluation was performed using a hold-out validation strategy (60% training, 20% validation, 20% testing). No external dataset was available to further validate the model, and no k-fold cross-validation was implemented due to computational constraints. For future work, applying k-fold cross-validation (e.g., 5-fold or 10-fold) and testing on independent crystallization datasets would provide stronger evidence of the model's generalization ability

2.4. Model Architecture

In this study, a transfer learning approach was applied using MobileNetV2 as the backbone architecture. The pre-trained MobileNetV2 model, trained on ImageNet, was used without its top classification layer (*include_top = False*), with an input size of $128 \times 128 \times 3$ pixels. The backbone layers were frozen to retain their feature extraction capabilities, while a new classification head was added and trained on the crystallization dataset. The classification head consisted of a Global Average Pooling layer, followed by two fully connected dense layers with 256 and 128 neurons respectively, both using ReLU activation. A final dense output layer with Softmax activation was used for binary classification (Nicotinamide and Ferulic Acid). The model was compiled with the Adam optimizer (learning rate = 0.00001) and categorical cross-entropy loss, with accuracy as the primary evaluation metric. The detailed architecture is summarized in Table 1.

Table 1. Detailed architecture of the CNN model with MobileNetV2 backbone and custom classification head

Layer Type	Detail	Activation	Notes
Input	$128 \times 128 \times 3$	-	Image input
MobileNetV2 Backbone	Pre-trained (ImageNet), frozen	ReLU	Feature extractor
Global Average Pooling	-	-	Dimension reduction
Dense	256 units	ReLU	Fully connected
Dense	128 units	ReLU	Fully connected
Output Dense	2 units	Softmax	Classification layer

2.4. Model Training and Testing

During the training stage, the CNN model described in the previous section MobileNetV2 as a frozen backbone combined with a custom classification head was used due to its ability to extract important visual features such as texture, shape, and patterns from images [22]. This extraction process takes place through convolutional layers, pooling layers, and fully connected layers, which together build a hierarchical representation of the data [23]. During training, internal parameters such as weights and biases are updated iteratively using optimization algorithms, including SGD and Adam, to minimize the loss function [24]. The primary goal is to achieve strong generalization ability, ensuring that the model not only performs well on the training data but also maintains good performance on unseen data [25], [26]. After the training process is completed, the model testing phase is carried out using a test dataset that is entirely separate from the training data, allowing the evaluation to objectively represent the model's performance under real-world conditions. In the context of crystallization image classification research, this stage is essential to assess how well CNN can recognize and accurately predict classes of pharmaceutical materials such as Nicotinamide and Ferulic Acid [27]. The test results are usually visualized in the form of a confusion matrix, which shows the distribution of correct and incorrect predictions for each category. This matrix is then used to calculate evaluation metrics such as accuracy, precision, recall, and F1-score [28]. Evaluation based on these metrics provides a comprehensive overview of the model's reliability while also identifying weaknesses that need improvement. Thus, the combination of training and testing stages forms an essential foundation for producing CNN models that are reliable, efficient, and applicable to image classification tasks, including those in the pharmaceutical field as well as in other domains [29].

2.5 Evaluate Model Performance

Model testing was carried out by evaluating performance using accuracy, which is calculated as the ratio between the number of correct predictions (both True Positives and True Negatives)

and the total number of observations. Accuracy measures the proportion of correct predictions against the entire sample and is calculated using Equation (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

The evaluation results are presented through a confusion matrix, a comprehensive analytical tool for comparing the model's predictions with the actual labels in the test data [30]. The confusion matrix facilitates the identification of the model's success in classifying each class and presents four key components: TP, TN, FP, and FN. These components can then be used to calculate precision, recall, F1-score, and accuracy metrics that are essential in assessing the performance of a classification model [31] [32].

Thus, the confusion matrix not only provides an overall picture of the model's effectiveness but also helps in understanding specific weaknesses that need to be addressed before further implementation [33]. Precision is used to measure how many cases predicted as positive by the model are truly positive. In crystallization image classification, precision evaluates the accuracy of the model in making positive predictions for each crystallization class. This metric provides insight into how precisely the model identifies a specific class [34], [35]. It is calculated using the following equation (2).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall is used to measure how well the model is able to identify all positive instances present in the actual data, which is important to ensure that image classifications such as crystallization images do not miss instances of the target class. This metric provides valuable insight into the model's sensitivity in recognizing each crystallization class [36]. The calculation of recall is based on the number of True Positives and False Negatives, as shown in Equation (3).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The F1-score is the harmonic mean between precision and recall, and it plays a crucial role in evaluating the performance of classification models, especially in cases where class distribution is imbalanced. In crystallization image classification, the F1-score provides a measure that balances the model's ability to make correct predictions (Precision) and to detect all true instances (Recall). This metric offers comprehensive insight into the model's performance, and its calculation is presented in Equation (4) as the harmonic mean of the two metrics [37].

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

3. RESULTS AND DISCUSSION

3.1. Model Training

The training data were fed into the model so that the network could comprehensively learn the patterns and characteristics of the images. The training process was carried out repeatedly through iterations known as epochs. In this study, training was run for 10 epochs, meaning the entire dataset was revisited ten times by the model. Each epoch provided the model with the opportunity to adjust weights and biases through optimization algorithms, allowing relevant patterns to become increasingly internalized. Through this repetition, the model gained a deeper understanding of the data, reduced the risk of misclassification, and improved its generalization ability on new data [38].

In the first iteration, the model achieved an accuracy of 83% with a computation time of 11 seconds, reflecting the initial stage of the model's learning process on the dataset. As the iterations increased, the model gradually showed improved accuracy due to continuous optimization of its parameters. By the 10th iteration, the model reached a much higher accuracy of 98%, with a shorter computation time of only 2 seconds. This indicates that the model not only learned better but also became more efficient in computation. The accompanying figure complements this analysis by presenting the training model's performance graph, which visualizes the trend of increasing accuracy and decreasing loss throughout the training process.

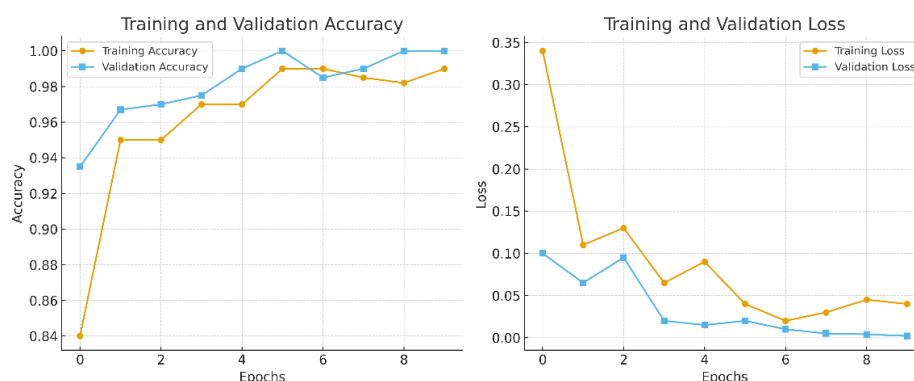


Figure 5 Training Results

Figure 5 shows that the training and validation accuracy increased sharply during the initial epochs and then quickly stabilized around 0.99–1.00. The loss curves also demonstrated a consistent decrease, with the validation loss appearing lower and more stable than the training loss. Although the results indicate that the model effectively recognized data patterns, the rapid convergence within only 10 epochs on a small dataset suggests a potential risk of overfitting. Therefore, the reported performance should be interpreted with caution, and future studies are encouraged to apply strategies such as early stopping, learning-rate scheduling, and larger or external datasets to ensure more reliable generalization.

3.2. Model Testing

After the training stage was completed, the next step was model testing, which aimed to evaluate the ability of the CNN to classify pharmaceutical material images. This testing process was carried out using test data that were never involved during the training stage, so that the evaluation results could objectively reflect the model's performance. In this study, the test data used amounted to 20% of the total dataset, consisting of 300 crystallization images divided into two classes: Nicotinamide and Ferulic Acid.

Feeding the test data into the model was intended to assess how well the network could generalize the patterns it had learned from the training data and apply them to new, unseen data. The testing results were then visualized using a confusion matrix, which displayed the distribution of correct and incorrect predictions for each class, thereby providing a more comprehensive overview of the model's classification performance. This visualization, presented in Figure 6, not only illustrates the evaluation results but also highlights the strengths and weaknesses of the model in distinguishing between different classes, offering deeper insights into its predictive capability.

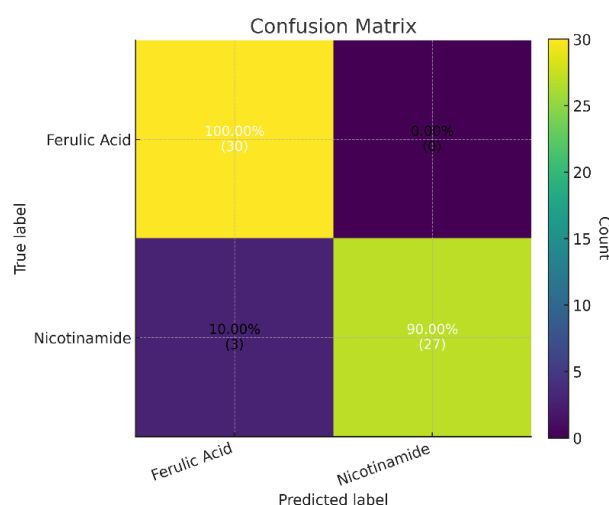


Figure 6. Confusion Matrix

Figure 6 presents the confusion matrix with explicit class labels. The figure indicates that for the Ferulic Acid class, the model predicted 30 images as belonging to Ferulic Acid, with no errors in classification. Meanwhile, for the Nicotinamide class, 27 images were correctly classified, while 3 images were misclassified as Ferulic Acid, which are considered False Positives (FP).

Table 2. Model Evaluation Matrix

Class	Precision	Recall	F1-Score	Class Accuracy
Ferulic	91%	100%	95%	100%
Nicotinamide	100%	90%	95%	90%
Overall	-	-	-	95%

Table 2 show that the model performed well in classifying the two classes. For the *Ferulic* class, it achieved 100% recall, 91% precision, and an F1-score of 95%, indicating that all *Ferulic* samples were correctly identified, although a few *Nicotinamide* samples (10%) were misclassified. For the *Nicotinamide* class, the model obtained 100% precision, 90% recall, and an F1-score of 95%, meaning all *Nicotinamide* predictions were correct, but three samples were misclassified as *Ferulic*. Overall, the model reached an accuracy of 95% with balanced performance across both classes. The slight recall limitation for *Nicotinamide* suggests that further improvements, such as expanding the training dataset or applying advanced data augmentation techniques, could enhance performance. These findings are consistent with recent studies that demonstrate the effectiveness of CNN-based approaches in medical image classification and highlight the need for continuous refinement to improve recall in minority classes.

These results are consistent with recent literature highlighting the effectiveness of CNN architectures in medical image classification. For example, [39], presented a comprehensive survey showing rapid progress in CNN design, including the use of hybrid architectures and hyperparameter optimization to improve performance in various medical classification tasks. Similarly [10], reported high accuracy in disease classification using X-ray, MRI, and ultrasound images, supported by advanced augmentation and transfer learning techniques. In the specific domain of urology, [40], achieved 98.8% sensitivity and 98.2% specificity in urinary cell classification by employing a fine-tuned VGG-16 CNN model. These findings reinforce that CNN-based methods are a promising and effective approach for early disease detection, including urinary protein analysis using dipstick tests. Although the proposed CNN model achieved a relatively high accuracy of 95%, this result should be interpreted with caution due to the limited dataset size (300 images across two classes). Small datasets increase the risk of overfitting and reduce the robustness of the evaluation. In practical terms, the lack of large-scale data may limit the generalizability of the model to new crystallization images. This limitation highlights the need for future studies to expand the dataset, either by collecting more experimental samples or by employing publicly available crystallization repositories.

4. CONCLUSION

This study demonstrates that the application of CNN can provide an effective solution for classifying crystallization images of pharmaceutical raw materials. The data collection process using hot-stage microscopy produced high-quality images of two types of materials, namely Nicotinamide and Ferulic Acid. Through preprocessing steps including normalization, resizing, and augmentation, the data became more homogeneous, enabling the CNN model to better recognize patterns. Training results showed an increase in accuracy from 83% in the initial epoch to 98% by the tenth epoch. Evaluation using the confusion matrix revealed that the model perfectly classified Ferulic Acid images (100%), while the Nicotinamide class still experienced a 10% misclassification rate. Nevertheless, the overall accuracy reached 95%, proving the effectiveness of CNN in recognizing complex crystallization patterns. The precision, recall, and F1-score also reflected consistent and reliable performance. The superiority of these results compared to previous studies highlights that CNN methods are not only relevant for medical image classification but also highly promising in the pharmaceutical field. These findings make a concrete contribution to the automation of drug material characterization processes, thereby improving efficiency, ensuring

quality consistency, and reducing reliance on manual evaluations prone to error. In conclusion, this study successfully achieved its main objective: developing a reliable CNN-based crystallization image classification system. Limitations remain in the Nicotinamide class, suggesting that future research should explore more complex network architectures, additional data augmentation techniques, or the integration of attention mechanisms to enhance model performance. In addition, the rapid convergence to near-perfect accuracy within only ten epochs, combined with the relatively small dataset size, may indicate a risk of overfitting. Future studies should therefore incorporate larger datasets, external validation, and training strategies such as early stopping or learning-rate scheduling to ensure more robust and generalizable results.

ACKNOWLEDGEMENTS (11 PT)

The authors would like to thank the Directorate of Research and Community Service, Ministry of Higher Education, Science, and Technology for funding support through the Master Thesis Research Grant (PTM) 2025, under Contract Number: 0419/C3/DT.05.00/2025

REFERENCES (11 PT)

- [1] F. Artusio, R. Contreras-Montoya, and J. A. Gavira, "Advances in Pharmaceutical Crystals: Control over Nucleation and Polymorphism," *Crystals*, vol. 14, no. 9, pp. 4–6, 2024, doi: 10.3390/cryst14090805.
- [2] A. Cote *et al.*, "Perspectives on the Current State, Challenges, and Opportunities in Pharmaceutical Crystallization Process Development," *Cryst. Growth Des.*, vol. 20, no. 12, pp. 7568–7581, 2020, doi: 10.1021/acs.cgd.0c00847.
- [3] I. Indra, R. Rahman, and R. Yulianti, "Karakterisasi Karbamazepin Hasil Rekristalisasi Berbagai Pelarut Organik dengan Metode Slow Evaporation," *J. Farm. Dan Ilmu Kefarmasian Indones.*, vol. 8, no. 3, p. 227, 2021, doi: 10.20473/jfiki.v8i32021.227-234.
- [4] M. Lambert *et al.*, "Modular microfluidic platform for solubility measurement, nucleation statistics and polymorph screening of active pharmaceutical ingredients: Irbesartan, Rimonabant, Aripiprazole and Sulfathiazole," *J. Cryst. Growth*, vol. 616, 2023, doi: 10.1016/j.jcrysgro.2023.127252.
- [5] T. Jiko, "Crystallization in Pharmaceutical Manufacturing : Optimizing Drug Stability and Bioavailability," pp. 1–2, doi: 10.35248/2157-7064.25.16.602.Citation.
- [6] A. Kumar, P. Singh, and A. Nanda, "Hot stage microscopy and its applications in pharmaceutical characterization," *Appl. Microsc.*, vol. 50, no. 1, 2020, doi: 10.1186/s42649-020-00032-9.
- [7] M. Mousavi and S. Hosseini, "A deep convolutional neural network approach using medical image classification," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02646-5.
- [8] O. N. F. King, K. E. Levik, J. Sandy, and M. Basham, "CHiMP: deep-learning tools trained on protein crystallization micrographs to enable automation of experiments," *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 80, no. Pt 10, pp. 744–764, 2024, doi: 10.1107/S2059798324009276.
- [9] A. Muis, S. Sunardi, and A. Yudhana, "Medical image classification of brain tumor using convolutional neural network algorithm," *J. Infotel*, vol. 15, no. 3, pp. 227–232, 2023, doi: 10.20895/infotel.v15i3.964.
- [10] H. Laçi, K. Sevrani, and S. Iqbal, "Deep learning approaches for classification tasks in medical X-ray, MRI, and ultrasound images: a scoping review," *BMC Med. Imaging*, vol. 25, no. 1, 2025, doi: 10.1186/s12880-025-01701-5.
- [11] A. Yilmaz, A. A. Demircali, S. Kocaman, and H. Uvet, "Comparison of Deep Learning and Traditional Machine Learning Techniques for Classification of Pap Smear Images," 2020, [Online]. Available: <http://arxiv.org/abs/2009.06366>
- [12] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognit. Lett.*, vol. 141, pp. 61–67, 2021, doi: 10.1016/j.patrec.2020.07.042.
- [13] Y. Kang *et al.*, "An Enhanced Deep Learning-Based Pharmaceutical Crystal Detection with

- Regional Filtering,” *Crystals*, vol. 14, no. 8, 2024, doi: 10.3390/cryst14080709.
- [14] J. Milne, C. Qian, D. Hargreaves, Y. Wang, and J. Wilson, “Not getting in too deep: A practical deep learning approach to routine crystallisation image classification,” *PLoS One*, vol. 18, no. 3 March, pp. 1–16, 2023, doi: 10.1371/journal.pone.0282562.
- [15] F. E. Pujiono, J. Ekowati, T. Amrillah, and D. Setyawan, “Ferulic Acid-Nicotinamide Cocrystal: Synthesis, Experimental, and Computation Study,” *Sci. Technol. Indones.*, vol. 10, no. 2, pp. 402–410, 2025, doi: 10.26554/sti.2025.10.2.402-410.
- [16] D. Characteristics, “Tropical Journal of Natural Product Research Enhancing the Pharmaceutical Properties of Ibuprofen through Spherical Agglomeration-Co-Crystallization with Nicotinamide : A Comprehensive Study of,” vol. 8, no. July, pp. 7715–7721, 2024.
- [17] R. S. Jebur, M. H. B. M. Zabil, D. A. Hammood, and L. K. Cheng, “A comprehensive review of image denoising in deep learning,” *Multimed. Tools Appl.*, vol. 83, no. 20, pp. 58181–58199, 2024, doi: 10.1007/s11042-023-17468-2.
- [18] I. Corley, C. Robinson, R. Dodhia, J. M. Lavista Ferres, and P. Najafirad, “Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 3162–3172, 2024, doi: 10.1109/CVPRW63382.2024.00322.
- [19] E. Jain, J. Neeraja, B. Banerjee, and P. Ghosh, “A Diagnostic Approach to Assess the Quality of Data Splitting in Machine Learning,” 2022, [Online]. Available: <http://arxiv.org/abs/2206.11721>
- [20] H. Bichri, A. Chergui, and M. Hain, “Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.
- [21] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, “Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images,” *Sci. Data*, vol. 9, no. 1, pp. 1–9, 2022, doi: 10.1038/s41597-022-01618-6.
- [22] S. Chatterjee and P. Zielinski, “On the Generalization Mystery in Deep Learning,” 2022, [Online]. Available: <http://arxiv.org/abs/2203.10036>
- [23] S. Jelassi and Y. Li, “Towards understanding how momentum improves generalization in deep learning,” *Proc. Mach. Learn. Res.*, vol. 162, pp. 9965–10040, 2022.
- [24] M. Lashkari and A. Gheibi, “Lipschitzness effect of a loss function on generalization performance of deep neural networks trained by Adam and AdamW optimizers,” *AUT J. Math. Comput.*, vol. 5, no. 4, pp. 361–375, 2024, doi: 10.22060/ajmc.2023.22182.1139.
- [25] L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin, “Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning,” *Proc. Mach. Learn. Res.*, vol. 235, pp. 62476–62509, 2024.
- [26] M. Wang and C. Ma, “Generalization Error Bounds for Deep Neural Networks Trained by SGD,” 2022, [Online]. Available: <http://arxiv.org/abs/2206.03299>
- [27] L. Zewen, L. Fan, Y. Wenjie, P. Shouheng, and Z. Jun, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE Trans. neural networks Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2021.
- [28] Y. L. Ma and M. Rho, “Towards the hadron–quark continuity via a topology change in compact stars,” *Prog. Part. Nucl. Phys.*, vol. 113, pp. 1–87, 2020, doi: 10.1016/j.ppnp.2020.103791.
- [29] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2022, doi: 10.1109/TPAMI.2021.3059968.
- [30] J. Görtler *et al.*, *Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels*, vol. 1, no. 1. Association for Computing Machinery, 2022. doi: 10.1145/3491102.3501823.
- [31] K. Riehl, M. Neunteufel, and M. Hemberg, “Hierarchical confusion matrix for classification

- performance evaluation,” *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 72, no. 5, pp. 1394–1412, 2023, doi: 10.1093/jrssc/qlad057.
- [32] A. Farhan AlShammari, “Implementation of Model Evaluation using Confusion Matrix in Python,” *Int. J. Comput. Appl.*, vol. 186, no. 50, pp. 975–8887, 2024.
- [33] S. Kanakala and S. Ningappa, “Detection and Classification of Diseases in Multi-Crop Leaves using LSTM and CNN Models,” *J. Innov. Image Process.*, vol. 7, no. 1, pp. 161–181, 2025, doi: 10.36548/jiip.2025.1.008.
- [34] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.
- [35] A. Tharwat, “Classification assessment methods,” *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
- [36] H. Kaur and D. Kaur Sandhu, “Evaluating the Effectiveness of the Proposed System Using F1 Score, Recall, Accuracy, Precision and Loss Metrics Compared to Prior Techniques,” *Int. J. Commun. Networks Inf. Secur.*, vol. 15, no. 4, pp. 368–383, 2023, [Online]. Available: <https://ijcnis.org>
- [37] M. Sitarz, “Extending F1 Metric, Probabilistic Approach,” *Adv. Artif. Intell. Mach. Learn.*, vol. 3, no. 2, pp. 1025–1038, 2023, doi: 10.54364/AAIML.2023.1161.
- [38] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.
- [39] F. A. Mohammed, K. K. Tune, B. G. Assefa, M. Jett, and S. Muhie, “Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature,” *Mach. Learn. Knowl. Extr.*, vol. 6, no. 1, pp. 699–735, 2024, doi: 10.3390/make6010033.
- [40] A. Teramoto, A. Michiba, Y. Kiriya, E. Sakurai, R. Shiroki, and T. Tsukamoto, “Automated Classification of Urinary Cells: Using Convolutional Neural Network Pre-trained on Lung Cells,” *Appl. Sci.*, vol. 13, no. 3, 2023, doi: 10.3390/app13031763.