2026 International Nuclear Atlantic Conference – INAC 2026
Escola de Guerra Naval, Rio de Janeiro, RJ, August 24 - 28, 2026

Evaluating CNN Robustness in Small Datasets: A Comparative Study on Two-Phase Flow Image Identification of Two-Phase Flow

S.M.C. Schott[1], M.C.B da Silva[2] and R.N. de Mesquita[3]

[1] *sandro.schott@usp.br,* Instituto de Pesquisas
Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242
marconeseng@gmail.com, Instituto de Pesquisas
Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242
[2] *rnavarro@ipen.br,* Instituto de Pesquisas
Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242

1. Introduction

The study of two-phase flow is essential across various applications, including nuclear power plants [1], where the identification of flow regimes is critical for ensuring safe and efficient operations [2]. A non-invasive approach to identifying these regimes involves flow imaging coupled with computational models for automated classification. Previously, classical machine learning techniques [3,4] were employed for this purpose; however, methods based on Deep Neural Networks (DNN) have recently been proposed [5].

The present work performs the classification of two-phase flow images using Convolutional Neural Networks (CNNs), expanding upon the research of previous studies [1][6]. In those prior works, classification was conducted using Fuzzy logic[1] and Self-Organizing Maps (SOM)[6] for images representing the midpoint of the three stages of chugging instability [7]. Consequently, those studies utilized highly characteristic images of each stage and a relatively small dataset of 96 images (32 per class).

Two aspects of these previous works are of significant interest, particularly those utilizing the SOM technique[6]. First, they perform classification using two variations of training and testing set ratios (50%/50% and 75%/25%), providing insights into the amount of data required for the model to achieve the target performance. Second, they introduce a protocol of 1000 training repetitions to calculate an average performance. This is relevant because model performance can vary across individual training runs, especially with small datasets where such variance is more pronounced, meaning a distribution better represents the models' true capability. These two methodological aspects are replicated in this work using CNNs to ensure a robust comparison. This issue was briefly addressed in [5][8], and the present study offers an opportunity to investigate it in depth.

Accordingly, this work aims to compare the predictive performance of modern CNNs with the classical techniques previously used for the same task, employing a methodology closely aligned with those earlier studies. Hence, this research extends the evaluation of CNN performance for two-phase flow classification using a significantly reduced dataset, contributing to the understanding of the minimum number of images per class required for effective classification.

2. Methodology

2.1 image acquisition

This study utilizes a dataset containing images of the three stages of chugging instability, acquired from the Natural Circulation Circuit (NCC) experimental setup. This same dataset was employed in previous studies [1][6]. While the complete database comprises approximately 1,500 images, the present research utilizes a subset of 96 images, consisting of representative samples from the midpoint of each stage. The images were partitioned into two distinct ratios for the training and testing subsets: 50%/50% and 75%/25%.

2.2 Models

This study expand upon previous research by implementing Convolutional Neural Networks (CNNs). Given the limited dataset of only 96 images, a consolidated and pre-trained architecture, VGG-16 [9], was selected. This CNN was pre-trained on the ImageNet [10] dataset, which contains millions of images. Fine-tuning was then performed on the current dataset by unfreezing and training the fully-connected layers. VGG-16 consists of 16 layers, with all convolutional layers utilizing 3x3 filters; this architecture demonstrated the significance of network depth in enhancing predictive capacity. In this work, the final convolutional layer was unfrozen, and a dense layer with 512 neurons was implemented. Additionally, a linear baseline model connected to each image pixel was utilized, serving as a critical benchmark for comparison with previous studies.

3. Results and Discussion

A Linear model (serving as a benchmark) and a CNN (VGG-16) were trained and evaluated 1000 times across two different training-to-testing set ratios: 50%/50% and 75%/25%. The accuracy results from the 1000 evaluations on the testing subset for the Linear model are shown in Figure 1.

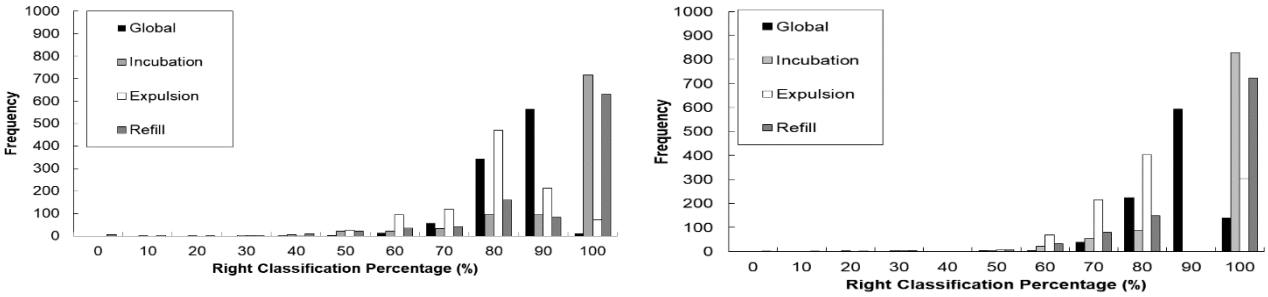
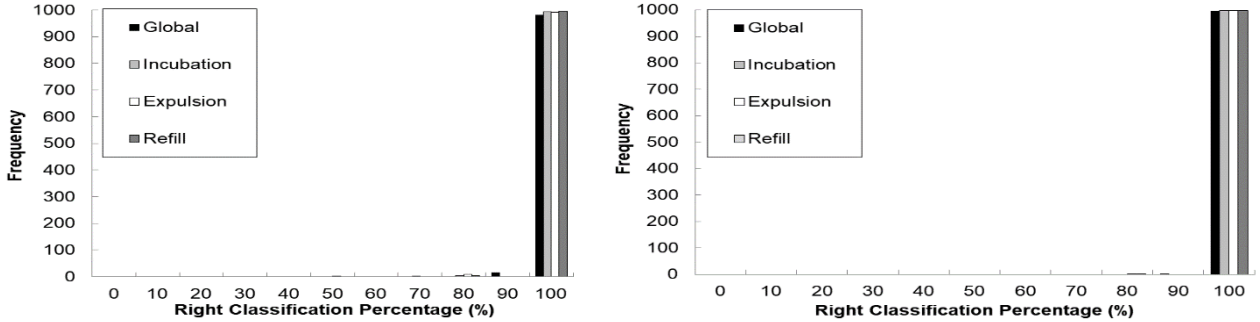


Figure 1: Accuracy distribution across 1,000 training runs of the linear model, using test subset proportions of 50% (left) and 25% (right).

As shown in Figure 1, the model's accuracy distribution ranges from 50% to 100%. This variation demonstrates that with a limited number of images, the Linear model exhibits significant variability in its results, justifying the implementation of a large number of training runs. The figure reveals high classification accuracy for the "Incubation" and "Refill" classes; however, performance is lower for the "Expulsion" class, where the most frequent errors involve misclassifying "Expulsion" as either "Refill" or "Incubation." The global mean accuracy was 89.96% and 93.98% for the 50% and 25% test ratios, respectively, indicating that an increase in the number of training images enhances predictive performance. Figure 2 presents the accuracy results for the CNN model (VGG-16).

Figure 2: Accuracy distribution across 1,000 training runs of the CNN model (VGG-16), using test subset proportions of 50% (left) and 25% (right).



As observed in Figure 2, the majority of the training runs for both ratios (50% and 25%) achieved an accuracy of 100%. Among of 1,000 runs, the accuracy fell below 100% only 18 and 3 times, respectively. The global mean accuracy was 99.90% for the 50% ratio and 99.99% for the 25% ratio. Consistent with previous observations, an increase in the number of training images leads to enhanced predictive performance. Based on these results, a comparison can be made with prior studies that utilized Fuzzy logic and SOM techniques; these values are presented in Table I.

Table I: Comparison with previous results for the 96-image classification task. Mean accuracy across 1,000 training runs.

	Test Subset Percentage		
	25%	50%	100%
Fuzzy[1]	-	-	90.63%
SOM[6]	93.98%	88.75%	-
Linear	92,19%	89,96%	-
CNN(VGG-16)	**99,99%**	**99,90%**	-

Table I enables a direct comparison with previous studies addressing the same task. It is observed that only the Fuzzy logic approach utilizes 100% of the data; nevertheless, this result serves as a valuable performance benchmark. The linear model demonstrates performance very similar to that of the SOM, being slightly inferior at the 25% test ratio and slightly superior at 50%. However, these models are fundamentally distinct: the SOM utilizes few parameters and operates under the unsupervised learning paradigm, whereas the Linear model, though conceptually simple and lacking non-linear elements, is computationally intensive. The linear model employs a high number of parameters (372,099), with each neuron directly connected to an individual pixel. In contrast, the VGG-based model achieves nearly 100% accuracy, failing to reach this performance in less than 1% of the training runs.

4. Conclusions

The classification of two-phase flow images was performed using CNNs on a limited dataset of only 96 images. The training process was repeated 1,000 times, and in over 99% of these iterations, the CNN achieved 100% accuracy, outperforming previous studies that utilized Fuzzy logic and SOM. Consequently, from a predictive performance standpoint, this specific problem is considered resolved through the application of CNNs. This represents a significant milestone for future applications of these models. However, regarding real-world scenarios, it is crucial to investigate how minor variations in the experimental setup during image acquisition affect model performance—studies referred to as domain shift. Such investigations will be conducted in future works.

Acknowledgements

This research was supported by Comissão Nacional de Energia Nuclear (CNEN).

References

[1] R. N. De Mesquita, et al. “Classification of natural circulation two-phase flow patterns using fuzzy inference on image analysis”. *Nuclear Engineering and Design*, p. 8, (2012).
[2] Ling, Chen, and Takashi Hibiki. "Interfacial structure-based flow regime characterization for gas-liquid two-phase flow." *International Journal of Heat and Mass Transfer* 263 (2026).

[3] C. Shanthi; N. Pappa, “An artificial intelligence based improved classification of two-phase flow patterns with feature extracted from acquired images”, *ISA Transactions,* (2017).
[4] G. Huang; et al. “Flow regime identification of mini-pipe gas-liquid two-phase flow based on textural feature series”. *IEEE International Instrumentation and Measurement Technology Conference* (2013).
[5] Schott, S. M. C., et al. "Convolutional neural network-based pattern recognition in natural circulation instability images." *Concilium*, (2024).
[6] R. N. De Mesquita, et al. “Classification of natural circulation two-phase flow image patterns based on self-organizing maps of full frame DCT coefficients”, *Nuclear Engineering and Design* (2018).
[7] J. A. Boure; A. E. Bergles; e L. S. Tong, “Review of two-phase flow instability”, *Nuclear Engineering and Design* (1973).
[8] Schott, S. M, C. “Classificação de padrões de escoamento bifásico por meio de redes neurais convolucionais*”*. Universidade de São Paulo, (2022).
[9]Simonyan, K.; Zisserman, A. “Very Deep Convolutional Networks for Large-Scale Image Recognition**”**.*3rd International Conference on Learning Representations*, ICLR 2015, San Diego, CA, USA, May 7-9, (2015)
[10] Deng, Jia et al. “Imagenet: A large-scale hierarchical image database*”. 2009 IEEE conference on computer vision and pattern recognition*, (2009)