Ensemble-Based CNN Approach for Gastric Cancer Classification in Histopathological Images

Camilly L. C. Silva*, Rodrigo M. S. Veras*, Rodrigo N. Borges*,
Ivan S. Silva*, Vinicius P. Machado* and François F. R. Barbosa†
*Universidade Federal do Piauí (UFPI), Teresina, Piauí, Brazil
Email: {camilly.costa, rveras, r_borges, ivan.silva, vinicius}@ufpi.edu.br
†Instituto Federal do Maranhão (IFMA), Timon, Maranhão, Brazil
Email: francois.barbosa@ifma.edu.br

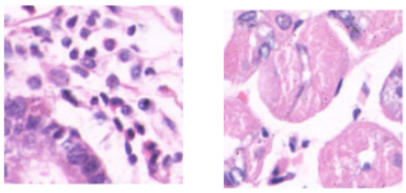
***Abstract*—Gastric cancer is one of the leading causes of cancer-related deaths worldwide and is often diagnosed at advanced stages. Histopathological analysis is essential for diagnosis but heavily relies on the specialist's expertise and is prone to human error. This study presents the development of a model based on Convolutional Neural Networks (CNNs) for identifying gastric cancer in histopathological images. The approach incorporates transfer learning, data augmentation, cross-validation, and an ensemble combining ResNet101, MobileNet, and EfficientNetB3. The model achieved an accuracy of 95.09% and a Kappa score of 89.72%. The results highlight the model's strong potential for practical application in supporting medical diagnosis.**

I. Introduction

Cancer is characterized by uncontrolled cell growth, forming malignant tumors that can invade nearby tissues and spread throughout the body [1]. Gastric cancer is one of the most prevalent types worldwide and is associated with high mortality, mainly because it usually shows no early symptoms, making diagnosis challenging [2], [3]. However, survival rates increase significantly when the disease is detected in its early stages [4], highlighting the importance of early and accurate diagnosis.

Recent advances in Computer-Aided Diagnosis (CAD) systems, particularly those using Convolutional Neural Networks (CNNs), have shown promising results in gastric cancer detection. These models can process large volumes of histopathological images efficiently, reducing subjectivity in exam interpretation and supporting medical decision-making [5].

In this work, we propose a CNN-based approach for classifying histopathological images of gastric tissue into healthy and cancerous samples. Figure 1 illustrates representative examples.



(a) (b)

Fig. 1. Example of gastric histopathological images. (a) non-cancerous and (b) cancerous.

The remainder of this paper is organized as follows: Section 2 reviews related work on AI-based histopathological analysis. Section 3 details the proposed methodology, including datasets, preprocessing steps, and evaluation metrics. Section 4 presents the experimental results, and Section 5 concludes with final remarks and future directions.

II. Related Work

In recent decades, significant advances have been achieved in the identification of gastric cancer in histopathological images through deep learning. Several approaches have been proposed, ranging from classical neural network models to more sophisticated methods.

Sharma et al. [6] employed CNNs and data augmentation to overcome the limitation of a small image dataset, achieving 96.82% accuracy with an adaptation of AlexNet.

Li et al. [7] proposed GastricNet, a customized deep learning architecture that reached 97.93% accuracy, outperforming well-known models such as AlexNet and ResNet-50.

Hu et al. [8] introduced the GasHisSDB dataset and demonstrated that CNNs such as VGG16 and ResNet50 (with accuracy around 96%) outperformed classical machine learning approaches and the Vision Transformer (ViT) model in the classification of the dataset images.

Yong et al. [9], using the same GasHisSDB dataset, combined transfer learning with an ensemble of CNNs. This approach achieved 99.20% accuracy, the best performance among the evaluated models, highlighting the potential of ensembles.

The reviewed studies demonstrate substantial progress in the application of deep learning techniques for gastric cancer detection, as well as the initial adoption of ensemble-based approaches aimed at improving the effectiveness of the methods. Despite these advances, challenges remain, particularly regarding the need for large annotated datasets and the difficulty of ensuring model generalization.

III. Methodology

In this study, we aimed to identify gastric cancer in histopathological slides using Python with the Keras and TensorFlow libraries. We describe the proposed model, the image dataset, the preprocessing and data augmentation tech-

niques, the neural networks employed, the validation set, the classification ensemble, and the evaluation metrics used to assess the effectiveness of the method.

A. Proposed Approach

We applied preprocessing and data augmentation techniques, such as rotations and flips, to standardize and expand the image dataset, thereby improving the model's generalization capability. The training was performed using five-fold cross-validation and shallow fine-tuning of pretrained models, with 65 epochs and predefined batch sizes. Several CNN architectures widely recognized in medical image analysis (MobileNet, ResNet50, MobileNetV2, ResNet101, VGG19, DenseNet121, and EfficientNetB3) were evaluated and adapted for detecting patterns in histopathological images related to gastric cancer identification.

From the five best-performing architectures, we constructed ensembles by combining them in groups of three, which resulted in a total of ten majority-voting committees. This strategy enhanced the robustness and accuracy of automated classification by leveraging complementary strengths of different CNNs. Finally, the ensembles were evaluated on different image sets, demonstrating both effectiveness and adaptability to new clinical scenarios.

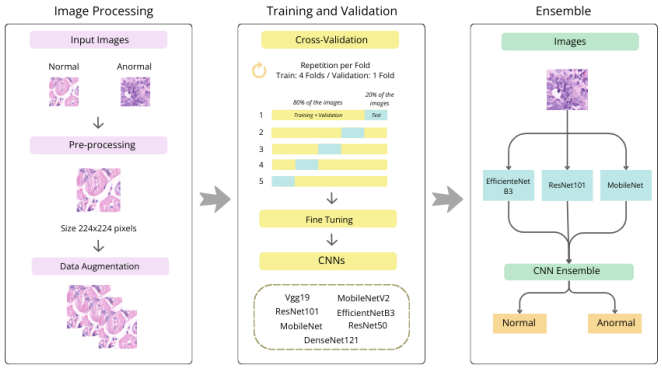


Fig. 2. Model Refinement and Training.

B. Image Dataset

The dataset used in this study, GasHisSDB, was developed through a collaboration between Longhua Hospital of Shanghai University of Traditional Chinese Medicine, Northeastern University, and Liaoning Cancer Hospital and Institute [8]. It contains 245,196 patches extracted from 600 whole-slide images (WSIs) with resolutions of 80×80, 120×120, and 160×160 pixels, divided into normal (148,120) and abnormal (97,076) samples.

In this work, we used only the 160×160 resolution, totaling 33,284 images (13,124 abnormal and 20,160 normal), balancing diagnostic relevance with computational feasibility. The use of other resolutions will be explored in future work.

C. Image Preprocessing and Data Augmentation

Since most models adopt the standard input size of 224×224 pixels, all images were resized to this dimension to ensure compatibility across training, validation, and testing.

Although the dataset contains a large number of samples, data augmentation is essential to increase variability and reduce the risk of overfitting [10]; [11]. The applied transformations included 20% zoom variation, horizontal flip, 20° rotation, horizontal/vertical shift up to 20%, shear transformation, and pixel filling.

These augmentations were applied only to the training and validation sets, preserving the integrity of the test set for model evaluation.

D. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning models designed to process images by extracting spatial features and identifying relevant patterns. They are widely applied in medical imaging, supporting diagnostic tasks [12].

In this study, we evaluated seven ImageNet-pretrained architectures [13]: MobileNet, ResNet50, MobileNetV2, ResNet101, VGG19, DenseNet121, and EfficientNetB3, all known for their strong performance in image classification benchmarks [14].

E. Transfer Learning and Cross-Validation

Transfer learning accelerates training by reusing prior knowledge, such as features learned from ImageNet [12]. We applied Shallow Fine-Tuning, adjusting only the final layers while keeping earlier convolutional layers frozen. This approach improves efficiency, reduces the need for large datasets, and enhances generalization [15].

Robustness was ensured through stratified 5-fold cross-validation [16]. At each iteration, 20% of the samples formed the test set, while the remaining 80% were divided into training (64%) and validation (16%). This strategy maintained class balance and enabled reliable evaluation of generalization on unseen data.

F. Ensemble

The ensemble strategy combines multiple CNNs to improve system performance by leveraging their complementary strengths. In this study, committees were formed by grouping three CNNs selected among the five best-performing architectures in the individual evaluation stage. The final prediction was obtained through majority voting, where each network casts a vote and the class with the highest count is chosen as output [16].

This approach mitigates the risk of relying on a single model, as the final decision integrates the diverse perspectives of different networks. By capturing distinct image features, the ensemble enhances accuracy, stability, and reliability, which is particularly relevant in medical image analysis tasks such as gastric cancer detection.

G. Evaluation Metrics

In this study, model performance was assessed using the confusion matrix, which summarizes correct and incorrect

classifications through true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) [17].

From these values, five metrics were computed: Accuracy (overall correctness), Precision (proportion of correctly identified positives), Recall (ability to retrieve all positive cases), and F1-Score (harmonic mean of Precision and Recall, indicating their balance).

The Kappa Index was also employed to measure the agreement between predictions and ground truth, accounting for chance agreement. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \times 100 \quad (1)$$

where p_o is the observed accuracy and p_e the expected accuracy by chance [17]. Values above 80% indicate strong agreement, 60–80% substantial, and below 40% weak.

IV. Results

In this section, the results obtained from the experiments with the proposed method on the selected dataset are presented. The experiments were carried out using data augmentation techniques, pretrained models, and the Shallow Fine-Tuning strategy, which allowed the networks to be adapted to the specific characteristics of the dataset.

Table I shows the performance metrics of the neural networks evaluated individually. ResNet101 stood out with the highest Kappa index (86.50±0.72%), indicating balanced performance across metrics, with high accuracy (93.82±0.34%) and recall (96.62±0.25%). DenseNet121 also achieved competitive results, with a recall value similar to ResNet101, although its other metrics were slightly lower. MobileNet

TABLE I
Evaluation Results of CNNs (Mean ± Standard Deviation in %)

Modelo	A(%)	P(%)	R(%)	F1-Score(%)	K(%)
ResNet101	**93,82 ± 0,34**	93,41 ± 0,41	96,62 ± 0,25	**94,99 ± 0,27**	**86,50 ± 0,72**
MobileNet	93,06 ± 0,38	**94,52** ± 0,49	93,99 ± 0,90	94,25 ± 0,34	85,49 ± 0,77
DenseNet121	92,17 ± 0,42	91,03 ± 0,43	96,60 ± 0,27	93,73 ± 0,33	83,34 ± 0,90
ResNet50	91,35 ± 0,54	88,82 ± 0,79	**98,07 ± 0,26**	93,21 ± 0,40	81,36 ± 1,22
EfficientNetB3	91,12 ± 0,41	93,78 ± 0,85	91,42 ± 1,22	92,57 ± 0,39	81,24 ± 0,81
VGG19	90,33 ± 0,49	92,23 ± 0,47	91,78 ± 0,67	92,00 ± 0,42	79,79 ± 1,02
MobileNetV2	89,08 ± 0,83	93,31 ± 0,58	88,30 ± 1,18	90,73 ± 0,74	77,46 ± 1,67

achieved the highest precision (94.52±0.49%) among the models, with overall consistent performance. ResNet50 obtained the highest absolute recall (98.07±0.26%), which is particularly useful in scenarios prioritizing the complete detection of positive cases; however, it showed lower precision (88.82±0.79%), indicating a higher number of false positives.

EfficientNetB3 demonstrated solid performance, with a Kappa value above the excellence threshold (81.24±0.81%) and well-balanced metrics. Finally, VGG19 and MobileNetV2 presented lower performance, with the smallest Kappa indices.

Overall, deeper models such as ResNet101 and DenseNet121 tended to provide a better balance across metrics, whereas lighter architectures, such as MobileNetV2, although computationally efficient, exhibited lower accuracy and weaker agreement with the ground truth.

To assess the effectiveness of the ensemble strategy and the impact of its compositions, ten combinations were generated from the five best individual models (ResNet101, MobileNet, DenseNet121, ResNet50, and EfficientNetB3), three of which are reported in Table II. This analysis highlighted the importance of variability in committee composition for the final performance.

TABLE II
Performance comparison of different ensembles (mean ± standard deviation in %).

Comitê	A(%)	P(%)	R(%)	F1-Score(%)	K(%)
Comitê 1	**95,09 ± 0,10**	93,83 ± 0,10	93,70 ± 0,10	93,77 ± 0,10	**89,72 ± 0,20**
Comitê 2	94,91 ± 0,21	**94,93 ± 0,21**	**94,91 ± 0,21**	**94,89 ± 0,21**	89,27 ± 0,44
Comitê 3	93,82 ± 0,22	93,92 ± 0,21	93,82 ± 0,22	93,77 ± 0,23	86,87 ± 0,49

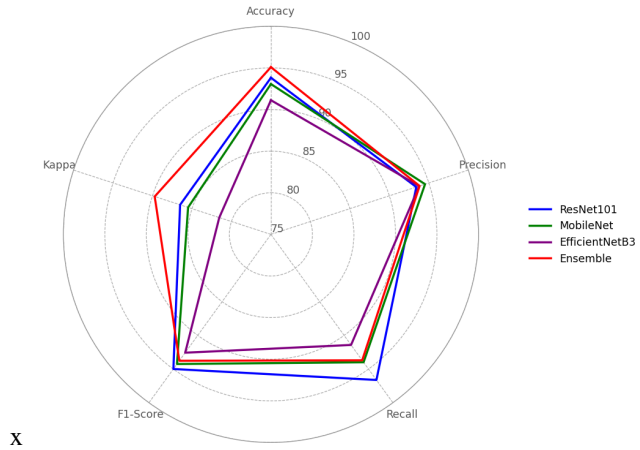
Committee 1 (ResNet101, MobileNet, and EfficientNetB3) achieved the best overall performance, with an accuracy of 95.09±0.10% and a Kappa index of 89.72±0.20%. Interestingly, this combination did not include DenseNet121 or ResNet50, which were among the models with the highest individual metrics, indicating that diversity and complementarity among CNNs can be more decisive than isolated performance.

Committee 2 (ResNet101, MobileNet, and DenseNet121) outperformed Committee 1 in Precision (94.93±0.21% vs. 93.83±0.10%), Recall (94.91±0.21% vs. 93.70±0.10%), and F1-Score (94.89±0.21% vs. 93.77±0.10%), suggesting that the choice of the most suitable model may depend on application-specific priorities. In contexts requiring greater emphasis on positive case detection (Recall) or on reducing false positives (Precision), Committee 2 may be more appropriate. Despite presenting a slightly lower Kappa index than Committee 1 (89.27% vs. 89.72%), its overall performance remained consistently high. In contrast, Committee 3 (DenseNet121, ResNet50, and EfficientNetB3) showed the lowest Kappa among the highlighted ensembles (86.87±0.49%). However, it is noteworthy that the Kappa of Committee 3 still surpassed that of ResNet101 (86.50%), which had been the best-performing individual model.

Overall, all evaluated committees achieved a Kappa index above 86% and other metrics (Accuracy, Precision, Recall, and F1-Score) exceeding 90%. These results reinforce that the ensemble technique consistently provides superior performance compared to individual models, helping to reduce variability across runs and promoting greater stability and reliability in diagnosis. Careful selection of ensemble components, considering their variability and complementarity, is therefore essential to optimize outcomes.

To illustrate the performance differences between Committee 1 and the three CNNs that compose it (ResNet101, MobileNet, and EfficientNetB3), Figure 3 presents a radar chart.

In this representation, the ensemble curve (red) appears as the outermost along the Accuracy axis, particularly in the Kappa index, which measures agreement beyond chance. This highlights the ensemble's ability to balance true positives and negatives. In contrast, Recall, F1-Score, and Precision are slightly lower than those of individual models such as ResNet101 (blue) and MobileNet (green), which reach higher peaks. Still, the ensemble sustains consistently high and well-



x

Fig. 3. Comparison of performance between the Ensemble model and its individual components.

distributed performance, supporting its adoption as the preferred strategy.

Figure 4 presents the training and validation accuracy and loss curves, showing effective learning without signs of *overfitting*. The similarity between the curves indicates good generalization to unseen data, supported by the use of *data augmentation* and shallow fine-tuning.

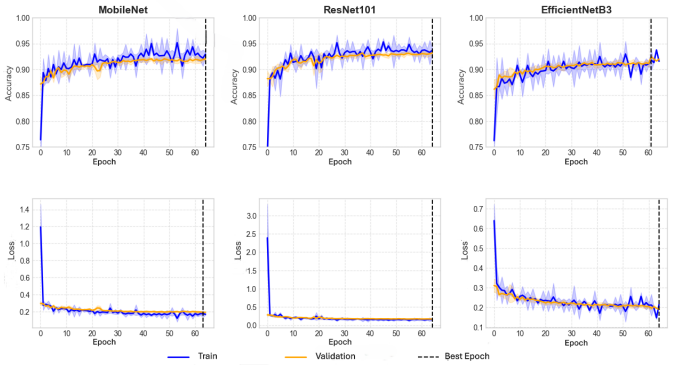


Fig. 4. Training and validation accuracy and loss curves for MobileNet, ResNet101, and EfficientNetB3.

The analysis of both individual models and ensembles, supported by training validation and result visualization, confirms the effectiveness of the proposed approach for gastric cancer classification. The best ensemble achieved 95.09% accuracy, surpassing all single models. In clinical contexts prioritizing Recall, Committee 2 proved superior (Recall: 94.91±0.21%, Precision: 94.93±0.21%, F1-Score: 94.89±0.21%). Thus, ensemble selection should align with specific clinical goals.

Although Yong et al. [9] reported higher accuracy (99.20%), differences in backbone models and fine-tuning strategies likely explain the gap. Still, the proposed method, combining shallow fine-tuning and dynamic ensembles, demonstrates robustness and clinical potential.

V. Conclusion

In this study, we proposed a CNN-based method with transfer learning and Shallow Fine Tuning for gastric cancer classification in histopathological images. The experimental results showed that the ensemble approach outperformed individual models, providing more stable and balanced performance. As future work, we plan to explore Deep Fine Tuning and expand the dataset to further improve accuracy and generalization.

References

[1] W. Ribeiro, O. Santiago, S. Oliveira, J. Souza, B. Fassarella, Y. Almeida, D. Jesus, R. Pontes, Q. França, M. Fassarella, L. Santos, and F. Neves, "Câncer de estômago: fatores de risco, prevenção e tratamento," *Brazilian Journal of Implantology and Health Sciences*, vol. 5, pp. 1098–1120, 2023.

[2] E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. Van Grieken, and F. Lordick, "Gastric cancer," *Lancet*, vol. 396, no. 10251, pp. 635–648, 2020.

[3] X. Fu, S. Liu, C. Li, and J. Sun, "Mclnet: A multidimensional convolutional lightweight network for gastric histopathology image classification," *Biomedical Signal Processing and Control*, vol. 80, p. 104319, 2023.

[4] M. Kato, T. Nishida, K. Yamamoto *et al.*, "Scheduled endoscopic surveillance controls secondary cancer after curative endoscopic resection for early gastric cancer: a multicentre retrospective cohort study by osaka university esd study group," *Gut*, vol. 62, pp. 1425–1432, 2013.

[5] Y. Zhao, B. Hu, Y. Wang, X. Yin, Y. Jiang, and X. Zhu, "Identification of gastric cancer with convolutional neural networks: a systematic review," *Multimedia Tools and Applications*, vol. 81, no. 8, pp. 11 717–11 736, 2022.

[6] H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, and P. Hufnagl, "Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," *Computerized Medical Imaging and Graphics*, 2017. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2017.06.001

[7] Y. Li, X. Li, X. Xie, and L. Shen, "Deep learning based gastric cancer identification," in *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, USA, 2018, pp. 182–185. [Online]. Available: https://doi.org/10.1109/ISBI.2018.8363550

[8] W. Hu, C. Li, X. Li, M. M. Rahaman, J. Ma, Y. Zhang, H. Chen, W. Liu, C. Sun, Y. Yao, H. Sun, and M. Grzegorzek, "Gashissdb: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer," *Computers in Biology and Medicine*, vol. 142, p. 105207, 2022.

[9] M. P. Yong, Y. C. Hum, K. W. Lai, Y. L. Lee, C. H. Goh, W. S. Yap, and Y. K. Tee, "Histopathological gastric cancer detection on gashissdb dataset using deep ensemble learning," *Diagnostics*, vol. 13, no. 1793, 2023. [Online]. Available: https://doi.org/10.3390/diagnostics13101793

[10] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[12] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[14] TensorFlow Developers, "tf.keras.applications: Models with pre-trained weights," https://www.tensorflow.org/api_docs/python/tf/keras/applications, 2024, acessado em 29 de junho de 2025.

[15] M. M. Taba, G. B. Junior, and R. M. de Souza, "A comparative study of shallow and deep fine-tuning for medical image classification," *Journal of Digital Imaging*, vol. 35, pp. 456–468, 2022.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.

[17] D. M. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation," School of Informatics and Engineering, Flinders University, Adelaide, Australia, Tech. Rep., 2007.