DeepHP: A New Gastric Mucosa Histopathology Dataset for Helicobacter pylori Infection Diagnosis

Wanderson Gonçalves e Gonçalves 1,2, Marcelo Henrique Paula dos Santos 1, Leonardo Miranda Brito 1, Helber Gonzales Almeida Palheta 1, Fábio Manoel França Lobato 3, Samia Demachki 1, Andrea Ribeiro-dos-Santos 1,2,*, and Gilderlanio Santana de Araújo 1, *

1 Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Graduate Program of Genetics and Molecular Biology, Federal University of Paraíba, Belém 6075-110, Brazil
2 Research Center on Oncology, Graduate Program of Oncology and Medical Science, Federal University of Paraíba, Belém 60703-000, Brazil
3 Laboratory of Applied Computing, Engineering and Geoscience Institute, Federal University of Western Paraíba, Santarém 68040-255, Brazil
*
Correspondence: akelyufpa@gmail.com (Â.R.-d.-S.); gilderlanio@gmail.com (G.S.d.A.)

Abstract: Emerging deep learning-based applications in precision medicine include computational histopathological analysis. However, there is a lack of the required training image datasets to generate classification and detection models. This phenomenon occurs mainly due to human factors that make it difficult to obtain well-annotated data. The present study provides a curated public collection of histopathological images (DeepHP) and a convolutional neural network model for diagnosing gastritis. Images from gastric biopsy histopathological exams were used to investigate the performance of the proposed model in detecting gastric mucosa with Helicobacter pylori infection. The DeepHP database comprises 394,926 histopathological images, of which 111 K were labeled as Helicobacter pylori positive and 283 K were Helicobacter pylori negative. We investigated the classification performance of three Convolutional Neural Network architectures. The models were tested and validated with two distinct image sets of 15% (59K patches) chosen randomly. The VGG16 architecture showed the best results with an Area Under the Curve of 0.998%. The results showed that CNN could be used to classify histopathological images from gastric mucosa with marked precision. Our model evidenced high potential and application in the computational pathology field.

Keywords: deep learning; Helicobacter pylori; computational pathology

1. Introduction

Gastric cancer is the fifth most common cancer worldwide and the fourth in deaths caused by cancer in 2020 [1]. Helicobacter pylori (HP) infection is the main risk factor accounting for around 89% of the distal gastric cancer cases around the world [2]. The HP is extremely adapted to the human gastric mucosa. However, disordered HP proliferation induces inflammation in the gastric mucosa, which can sequentially result in gastric cancer development [3]. It can bind to epithelial cells and prevent the immune response to cancer. HP has a high prevalence worldwide, of almost 50% [4], and 1–3% of the cases of H. pylori infection progress to gastric cancer [5].

Endoscopy is the main procedure for assessing HP infection and gastric cancer, followed by an histopathological biopsy analysis [6]. The histopathological analysis allows for identifying HP ubiquity and morphological alterations in the gastric mucosa. A gastric biopsy analysis is a highly time-consuming task. It can be affected by biotechnological factors such as staining techniques, errors in gathering biopsy sites, and also the pathologists’ subjectivity/experience [7,8].

Intelligent computational models are promising in the medical domain in terms of assisting clinical decisions [9]. Few studies have been proposed for gastric-related dis-
eases [10]. Currently, deep learning methods, such as Convolutional Neural Networks (CNNs), are the cutting-edge in research applications for gastric cancer analysis. Moreover, one of the main problems in deep learning applications is the scarcity of large image collections for training to generate classification, segmentation, or detection models, mainly due to complexities in obtaining well-annotated data [10,11]. The absence of high-quality datasets impairs the development of new and optimized tools for the biopsy image analysis of gastric mucosa diseases, making it nearly impossible to develop comparative studies between computational models. In addition, we previously identified several issues regarding model reproducibility caused by a lack of experimental setup information, such as uninformed network hyper-parameters, computational cost, and the availability of private or public image collections [10,12].

Our main contributions are summarized as follows. First, we provide DeepHP, a publicly curated dataset of histopathological images for HP classification for diagnostic studies. This dataset was built by following a detailed protocol to provide trustworthiness in the subjacency analysis. Among the 394,926 patched images, 283,921 were identified as negative for HP infection and 111,005 as positive for HP infection. Secondly, we performed extensive experiments and adopted transfer learning and fine-tuning strategies to analyze the performance of three CNN models in classifying HP infection with our dataset. The best CNN model showed an Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), F1-Score, and Mathews Correlation Coefficient (MCC) of 0.99, 0.96, and 0.95, respectively. The transfer learning and fine-tuning approach to histopathological images showed promising results and can be used for advancements in patient-level classification for diagnostic support.

Related Works

There are few research studies concerning the application of deep learning for classification, detection, or segmentation tasks to detect gastritis or HP infection. The challenge is to obtain reliable, robust, and curated histopathological image collections for training, testing, and validating CNN models. Beyond that, deep learning methods have shown encouraging results in the classification and detection of gastritis [13–15]. Those classification and detection models have been built using X-ray images, which are not invasive, and the applications showed good performance. In some ways, the automated analysis of endoscopic images were superior to that of endoscopists in the classification task [13–15].

Endoscopy images are widely used to detect HP infection with pre-trained networks (e.g., GoogleNet, ResNet-50, or Inception v3) based on transfer learning and fine-tuning strategies [16–21]. Interestingly, Nakashima et al. [22] trained a distinct CNN to classify endoscopy video frames between HP positive and negative samples.

Martin et al. [23] trained the HALO-AI software to classify gastric images with HP–related gastritis, reactive gastropathy, and histopathologically normal gastric mucosa. Their dataset consisted of 300 images of classic cases that best represented the desired pathology, divided into 100 for each category, and 106 other examples of non-classical gastric biopsies. Around 70% of the 300 images were used for training, and the other 30%, plus the 106 non-classical case images, were used for testing.

Despite several works dedicated to analyzing gastroenterological issues, we did not find public links for image datasets, including the previous research studies discussed in this section. Thus, our work contrasts by providing enough information and data to improve research replicability. The dataset with curated histopathological images labeled HP positive/negative is publicly available.

The remainder of this paper is organized as follows. Results are presented and discussed in Section 2. In Section 3, the material and methods are presented. A critical comparison of the obtained results with state-of-the-art is given in Section 2.3. Finally, the conclusions are presented in Section 4.
2. Results and Discussion

In this Section, we present and discuss the results of both generating the DeepHP and the transfer learning experiments.

2.1. The Working DeepHP Dataset

Deep learning has been applied to gastroenterological problems on private image collections from distinct research centers. Consequently, we did not find publicly available datasets, impairing the experiments’ transparency and reproducibility [10]. The growth of deep learning applications requires a large volume of reliable data to fill the gap in medical applications such as gastroscopy, colonoscopy, and digital pathology to properly train deep learning networks [24].

Our data were obtained from a pathological anatomy archive with diagnoses already established for each slide. The preprocessing strategy was performed to separate tissue and tissue-free images in patches. The filtering strategy resulted in a dataset with 283,921 patches diagnosed as \textit{H. pylori} negative and 111,005 patches diagnosed as \textit{H. pylori} positive. The agreement between the raters produced a $\hat{\kappa}$ coefficient of close to 0.95, which was evaluated as almost perfect. The overall accuracy of the filter defined as Rater A was 0.98, while Rater B achieved 0.97.

This dataset was used in the pre-trained CNN models to assess image classification, i.e., whether they were HP positive or negative. In our experiments, 70% of gastric slices (both \textit{H. pylori} positive and \textit{H. pylori} negative) were randomly selected for CNN training, 15% of slices were used for validation, and 15% were used for the testing steps.

2.2. Transfer Learning and Fine-Tuning Results

There is still a great deal of discussion about the number of images needed for deep learning. However, within the learning scenario, transfer learning is the most popular due to the versatility of using external data [25]. Thus, with this image collection, we used transfer learning techniques and fine-tuning to evaluate the performance of the following distinct CNN architectures: VGG16, InceptionV3, and ResNet50.

In general, studies that address deep learning applications in image classification, detection, or segmentation usually use accuracy as the predominant metric [10]. In addition, we evaluate CNN models with the F1-Score and MCC for an informative and truthful score in evaluating binary classifications [26]. The F1 score is an evaluation metric that considers recall and precision in its formulation; it is a more balanced metric than the unit elements of the confusion matrix. However, the result can be biased because it does not include the true negatives. Thus, the F1 score is easily influenced by the class labeled as positive [27]. In contrast, the MCC considers all cells from the confusion matrix and will only perform well if the classifications of the positive and negative classes are also equally good [26]. Therefore, the interpretation of these metrics depends on the study domain. The F1-Score will address problems that give preference to a specific positive class; however, if the cost of low precision and low recall is unknown, the MCC is preferable to the F1-score. We also performed AUC analyses to verify the sensitivity and specificity of the models.

The performance of each model trained without transfer learning is shown in Table 1. The VGG16 has significantly better performance than the InceptionV3 and ResNet50 models, according to MCC $\geq$ 0.88. Additionally, we evaluated the models by implementing the fine-tuning approach with the test data. Results for the fine-tuning performance are shown in Table 2. VGG16 outperforms InceptionV3 and ResNet50 after fine-tuning. The results in Tables 1 and 2 make clear the effectiveness of the fine-tuning approach; we observed an improvement in the pre-trained VGG16 and InceptionV3 models with the fine-tuning approach, while the ResNet50 model showed a decreased performance, which may be due to the new data domain or the choice of unfrozen blocks during the fine-tuning process.

VGG may perform equivalently in similar independent datasets. For example, HALO-AI implements a fully convolutional version of the VGG architecture [23]. Indeed, model
performance could be increased with hyperparameter optimization. Although the parameters for VGG16 are higher in this work, we still see that there is a good performance in the other models (ResNet50 and InceptionV3). The complexity of the models, the pre-training set, the adjusted hyperparameters, and the dataset configuration itself are factors that greatly influence the final result. Consequently, changing these factors can easily affect the performance of the CNN models.

Table 1. An overview of the evaluation metric results was obtained for the VGG16, InceptionV3, and ResNet50 models using the transfer learning strategy.

| Model    | Accuracy | Precision | Recall | Specificity | F1 Score | MCC  |
|----------|----------|-----------|--------|-------------|----------|------|
| VGG16    | 0.95     | 0.93      | 0.90   | 0.97        | 0.91     | 0.88 |
| InceptionV3 | 0.94     | 0.90      | 0.86   | 0.96        | 0.87     | 0.84 |
| ResNet50 | 0.95     | 0.85      | 0.98   | 0.93        | 0.91     | 0.87 |

Table 2. An overview of evaluation metric results was obtained for VGG16, InceptionV3, and ResNet50 using the fine-tuning approach.

| Model    | Accuracy | Precision | Recall | Specificity | F1 Score | MCC  |
|----------|----------|-----------|--------|-------------|----------|------|
| VGG16    | 0.98     | 0.94      | 0.99   | 0.98        | 0.96     | 0.95 |
| InceptionV3 | 0.96     | 0.88      | 0.98   | 0.95        | 0.92     | 0.90 |
| ResNet50 | 0.94     | 0.82      | 0.99   | 0.91        | 0.89     | 0.86 |

Additionally, we used the Receiver Operating Characteristic Curve (ROC curve) to determine the accuracy of the CNN decision. The ROC curve is derived from the graph of sensitivity against specificity, for which we calculated the AUC. No cut-off values (values that differentiate between positive and negative results) were used, and the ROC curve was obtained when plotting the values. Values ranged from 0 to 1, with values close to 1 indicating a high classification accuracy. The results of the ROC curves and the respective AUCs are shown in Figure 1. The ROC graph plots the false positive rate (FPR) against the true positive rate (TPR). Similarly, the Precision–Recall curve (PRC) is an alternative to the ROC curve for tasks with large deviations in the class distribution. Furthermore, PR curves can expose differences between algorithms that are not apparent in the ROC [28]. The results of the Precision–Recall curves and the respective AUCs are shown in Figure 1 which shows VGG16 has good performance (AUC = 0.997).

However, common knowledge in the scientific community points to unbalanced data as those that correspond to a dataset with significant and even extreme imbalances, which is not the case in this work. These imbalances are commonly attributed to class ratios in the order of 100:1, 1000:1, and 10,000:1, where in each case one class severely represents another [29,30].
2.3. Related Studies and Analysis

Our study surpassed others involving endoscopy imaging [17,18]. Deep learning analyses of HP infection by upper gastrointestinal endoscopy images showed a sensitivity and specificity of 86.7% and 86.7%, respectively, and the AUC equals 0.956 [18]. In [17], which implements GoogleNet fine-tuning to diagnose HP infection, provided a promising diagnostic model with an Area Under the Curve 0.96 and 0.95 using Bright Blue (BLI-bright) and linked color image laser image data.

In Martin et al. [23], the authors used HALO-AI image analysis on tissue with HP-related gastritis, reactive gastropathy, and gastric mucosa. Their dataset consisted of 300–100 images for each category and 106 other examples of non-classical gastric biopsies. Around 70% of the 300 images were used for training, and the other 30%, plus the 106 non-classical case images, were used for testing. The AUCs were 91.9% for normal, 100% for \textit{H. pylori}, and 94.0% for reactive gastropathy. The sensitivity and specificity were as follows: normal (73.7%, 79.6%), \textit{H. pylori} (95.7%, 100%), reactive gastropathy (100%, 62.5%) in 130 additional gastric biopsies. The HALO-AI software assigned color labels to the test biopsies, which correspond to the area of tissue assigned to a diagnosis by the DL algorithm, called the area distribution. For the results presented above, there was an area distribution cutoff of 40%.

The mentioned study used established software to train and test its dataset. This software allows the analyst to contour the areas of interest on the whole slide. Pathologists make this markup in the software itself, and this information is provided as input for
network training. Slides were taken at ×5.5 magnification. In addition, the authors used a cut-off in their evaluation metrics.

In comparison to our study, there are significant differences in the dataset acquisition method and image magnification. We used image patches, pre-trained CNN models with classification layer adjustments, and the application of fine-tuning with ×20 magnification. We obtained a classification result for HP infection and without HP infection, without cutoffs, with an AUC of 0.998. This result represents a discriminative power of identifying inflammatory profiles associated with HP infection at the level of image patches. The use of cutoffs can further enhance the ability of classification.

In recent years, the regular use of AI-based medical devices, including CNN applications, has received much attention. The FDA has already approved two medical devices, the IDX-DR for image analysis for the detection of diabetic retinopathy and one called Viz.AI, which analyzes images for indicators of stroke [31]. The use of these devices is becoming a reality in the medical context. That is why regulatory agencies are concerned about methodologies and these applications, which can increase the reliability and effectiveness of medical assistance. There are still discussions about how these artificial intelligence algorithms create their problem-solving rules. These applications become so complex that they hide what happens in multiple layers, called “black-box” algorithms. In Europe, the European Union’s General Data Protection Regulation (GDPR) implies that language needs to be fully explained; therefore, an algorithm cannot be used without revealing its “black-box”. However, given the robust results of software AI applications in the medical field, such as in the detection of melanoma, where CNN obtained results superior to a group of human experts [32], the FDA did not establish such a barrier and published a discussion paper that proposes a regulatory framework for software modifications based on artificial intelligence in computer-assisted diagnosis [31,33].

There are some limitations in our study that should be pointed out. First, we used a few examples to generate the database. This limitation makes the CNN constrained to the generated dataset. More images can help to provide a more significant dataset variability and, consequently, better generalization of the model. This limitation can be solved in future studies by including more slides. Obtaining health data is always a sensitive topic, with problems ranging from ethical and patient privacy issues to the selection of appropriate equipment for image acquisition. Second, training and testing data collection was performed at a single center. The use of data from other research/diagnostic centers could improve sample heterogeneity for model validation and generalization. However, as mentioned earlier, data acquisition in healthcare can be a difficult and time-consuming process. We seek cooperative agreements with other research centers to overcome this gap.

3. Material and Methods

This section is divided according to our outcomes, the DeepHP dataset, and the experiments with CNNs.

3.1. DeepHP Dataset

This subsection encompasses data acquisition, pre-processing, filtering, and validation.

3.1.1. Images Collection

The images used to build the dataset were collected from gastric biopsies stained with hematoxylin-eosin (H&E) using histopathological Whole Slide Imaging (WSI). All digital images are RGB, 0.16 µm per pixel, scanned with a ZEISS Microscope—Axio—imager.M2, ×20 magnification. Histopathological samples are anonymous images of human gastric mucosa embedded in paraffin fixed in formalin and stained with H&E from the pathological anatomy archive of Hospital Universitário João de Barros Barreto, Federal University of Pará (Belém, Pará, Brazil).

The original dataset consists of 13,921 images generated from 19 histopathological WSI, 14 gastric mucosae without morphological changes (9926 negative HP samples images),
and five with HP infection (3995 positive sample images). Each image has dimensions
of 2776 × 2080 pixels. The original DeepHP image collection can be requested by following
the instructions in the DeepHP section at https://www.lghm.com.br/datasets (accessed on
22 November 2022).

3.1.2. Image Pre-Processing

Each image (2776 × 2080 pixels) was split into smaller non-overlapped image patches
of 256 × 256 pixels using grid sampling (see a random sample in Figure 2). After this
process, our dataset comprised 1,113,680 patches sized to 256 × 256 pixels (794,080 negative
H. pylori and 319,600 positive H. pylori).

Figure 2. A random example of the original H&E image (2776 × 2080 pixels) and one patch sized to
256 × 256 pixels.

During the capture process of biopsy images, we observed the absence of gastric tissue
in some images after automatic patch cropping. We implemented a pre-processing strategy
to detect the tissue and the bottom of the slide. The goal was to keep samples filled with
at least 50% of gastric tissue in the patch. The images with less than 50% of observable
tissue in the patch were moved out of the working dataset. The approach is described in
Algorithm 1.

Algorithm 1: Image selection algorithm.

| Data: Collection of cropped histological images. |
| --- |
| Result: Curated histological image collection. |

1. Initialize ImageCollection;
2. foreach Image do
3. 
4. $I_{gs} \leftarrow \text{Image.toGrayScale}();$
5. $I_{mask} \leftarrow I_{gs}.\text{ApplyMedianMask}();$
6. $I_{wp} \leftarrow I_{mask}.\text{Binarize}();$
7. $wp \leftarrow I_{wp}.\text{GetWhitePixelProportion}();$
8. if $wp \leq 0.5$ then
   9. ImageCollection.add(Image);

The patches are RGB figures, which were transformed into grayscale. This step is
crucial to increasing the speed of image operations. Then, we apply a median mask
to reduce image noise. After this smoothing, the patches were linearized. Finally, the
proportion of white pixels is measured, which corresponds to the image background. If
this proportion is more than 50% (our percentage threshold), the patch is removed from
the working database; otherwise, it is retained. This process can be seen in Figure 3.
For filter validation, the patches were sampled according to Expression (1) as follows, with a 95% confidence interval:

\[
ss = \frac{z^2 p(1-p)}{\varepsilon^2} \left( \frac{1}{1 + \left( \frac{z^2 p(1-p)}{\varepsilon^2 N} \right)} \right)
\]

where \(ss\) is a sample size, \(N\) is the population size, \(\varepsilon\) corresponds to the margin of error (percentage in decimal format), \(p\) represents the expected proportion and \(z\) is the \(z\)-score, which is the number of standard deviations between a given ratio and the mean.

Subsequently, 400 samples were selected randomly for each of the 19 patch folders. Image binarization limits and white ratio were applied, as shown above. Two independent raters defined the filter thresholds in the sample data (see Table 3). Each rater received a copy of the same sample set, then the values were assessed through subjective inspection.

Afterward, the threshold average was applied to the original sample data patch, and each inspector evaluated the result independently. The result of the average thresholds between the evaluators is presented in Section 2.1.

### Table 3. Thresholds used in the separation filter for the sample data.

| Slide | Rater X Threshold | Rater Y Threshold | Average Threshold |
|-------|-------------------|-------------------|-------------------|
| 1     | 196               | 210               | 203               |
| 2     | 186               | 200               | 193               |
| 3     | 199               | 215               | 207               |
| 4     | 176               | 193               | 184.5             |
| 5     | 186               | 195               | 190.5             |
| 6     | 161               | 195               | 178               |
| 7     | 217               | 235               | 226               |
| 8     | 204               | 230               | 217               |
| 9     | 196               | 220               | 208               |
| 10    | 194               | 205               | 199.5             |
| 11    | 217               | 240               | 228.5             |
| 12    | 204               | 235               | 219.5             |
| 13    | 178               | 205               | 191.5             |
| 14    | 178               | 195               | 186.5             |
| 15    | 204               | 235               | 219.5             |
| 16    | 153               | 190               | 171.5             |
| 17    | 178               | 205               | 191.5             |
| 18    | 191               | 200               | 195.5             |
| 19    | 196               | 220               | 208               |

#### 3.1.3. Image Validation

The statistical concordance (with or without tissue) was made using the Kappa coefficient (\(\hat{\kappa}\)). The evaluation of the automatic fragment separation filter was carried out using
the confusion matrix and global accuracy. The global accuracy was calculated according to the ratio between the sum of true positives and true negatives \((TP + TN)\) and the sum of total positives and negatives \((P + N)\), as described in Equation (2).

\[
\text{accuracy} = \frac{TP + TN}{P + N}
\]  

(2)

Cohen begins by defining a confusion matrix \(k\) by \(k\) and calculates the \(\hat{\kappa}\) coefficient according to Equation (3) [34]):

\[
\hat{\kappa} = \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e}
\]  

(3)

Being the overall proportion of observed agreement \((\hat{p}_o)\):

\[
\hat{p}_o = \sum_{i=1}^{k} \frac{p_{ii}}{n},
\]  

(4)

and the overall proportion of chance-expected agreement \((\hat{p}_e)\):

\[
\hat{p}_e = \sum_{i=1}^{k} \frac{p_{i}p_{i}}{n^2}.
\]  

(5)

where \(k\) represents the evaluation categories, \(p_{ij}\) represents the number of samples evaluated by rater \(Y\) in category \(i\), \(p_{i}\) represents the number of sample elements evaluated by rater \(X\) in category \(i\) and \(n\) represents the total samples evaluated.

The Kappa coefficient \((\kappa)\) is suitable for a better model/annotation process assessment since it takes into account all the elements of the confusion matrix, different from the overall accuracy that uses only the main diagonal [35].

In the present work, the Kappa coefficient is compared to the indices proposed by Landis and Koch, 1977 [36], which characterize different ranges of values for kappa concerning the degree of agreement. The indices are shown in Table 4.

**Table 4.** Classification quality based on the Kappa coefficient values.

| Value of Kappa | Level of Agreement |
|---------------|--------------------|
| <0.00         | Poor               |
| 0.00–0.20     | Slight             |
| 0.21–0.40     | Fair               |
| 0.41–0.60     | Moderate            |
| 0.61–0.80     | Substantial         |
| 0.81–1.00     | Almost Perfect      |

3.2. Convolutional Neural Network

Convolutional Neural Networks (CNNs) are bio-inspired neural networks that learn through the spatial structure of the images and the dependence between pixels close to each other [37]. Its layers are divided into input, feature extraction, activation, classification, and output layers.

The CNN was implemented as a multilayer architecture based on artificial neural networks (ARNs) and the backpropagation algorithm. Computational costs have limited the development of this machine learning subfield, the amount of data available, and restrictions of classical techniques [38]. Figure 4 illustrates the structure of a CNN. The first layer in the process is the input layer. This step is responsible for placing the original images of the problem to be solved in the network. The next step is to insert feature extraction layers, which correspond to Convolutional and Grouping Layers (or subsampling), using kernels or spatial filters, extracting important features, and removing those not crucial to the process.
The activation function performs nonlinear transformation in the input data, which facilitates the learning and performance of more complex tasks. Each layer becomes an expert in identifying essential image features, such as colors, outlines, curves, formats, and others that depend on the application domain.

The convolutional kernel values are learned by the network in the training phase [39]. This process generates the known activation maps produced by each layer and assists in the feature extraction process. Since pool filters have no defined values, they are used as a logical operation to reduce the network’s processing load. These filters produce another image with a smaller dimension than the previous one, which can be the extraction of the maximum, medium, or minimum value of a pixel within the filter that will be applied to the image.

The image’s dimensions are reduced within the network. The output image is transformed into a 1D characteristic vector, a flat 1D layer at the end of the convolutional and pooling layers application. This long input data vector passes through the artificial neural network, usually a “Multilayer Perceptron” (MLP), to perform the feature classification.

In this way, a CNN can adjust the image dataset by reducing parameters and reusing weights, enabling the successful capture of the spatial and temporal dependencies of the image by applying relevant filters [40]. Based on the CNN model, we investigated the performance of the following three main CNN architectures: VGG16, InceptionV3, and ResNet50.

- **VGG16**: The first work that popularized CNNs in computer vision is VGG-Nets [41], which directly incorporated classic CNN architectures, and won first and second places in the location and ranking tasks in the ImageNet Challenge 2014, respectively. We opted for the VGG network with 16 weight layers (VGG16), namely thirteen convolutional layers and three fully-connected layers with a final softmax classifier. The convolution operations on the thirteen layers were performed using a \( 3 \times 3 \) pixel dimension kernel, with the \( W \) and \( b \) parameters being slipped over the \( \times \) pixels of each image resulting in a \( y \) output (Equation (6)). Kernel offset is pass-dependent and can be pixel-by-pixel or can skip multiple pixels. Convolutional layers act as feature extractors from the input images. The deeper the layers, the more specific the details that are extracted, and the earlier the layers, the more general the features that are extracted. The activation function of each convolutional layer is the Linear Rectified Unit (ReLU) for removing the linearity factor. The convolution operations results, also called feature maps, receive a pooling layer to reduce dimensionality, and in this case, max-pooling is applied in a \( 2 \times 2 \) pixel window, with step 2. Finally, three layers that are Fully-Connected (FC) follow a stack of convolutional layers, two layers have 4096 neurons, and the third performs 1000 classifications and therefore contains 1000 neurons. The final layer receives the soft-max function [42].

\[
y = f(Wx + b)
\] (6)
• INCEPTION V3: Proposed by Szegedy et al. in 2015, InceptionV3 uses initial blocks based on a typical CNN architecture [41]. This architecture achieved good performance with relatively low computational cost, which was enhanced by the orderly addition of RMSProp, label smoothing, $7 \times 7$ factoring, and BN auxiliary layers in the InceptionV2 network [43]. The network includes eleven Inception modules of five types, concatenated to obtain maximum resource extraction. Each module is branched and applies different kernel sizes ($1 \times 1$, $3 \times 3$, $5 \times 5$, and $7 \times 7$). These filters extract and concatenate different scales from feature maps. In the following step, $1 \times 1$ convolutions are applied for dimensionality reduction before computationally more expensive $3 \times 3$ and $5 \times 5$ convolutions. The factoring strategy of $5 \times 5$, and $7 \times 7$ convolutions into smaller convolutions ($3 \times 3$) or asymmetric convolutions ($1 \times 7, 7 \times 1$) is applied to reduce the number of parameters [44]. In summary, InceptionV3 uses symmetrical and asymmetrical components, including convolutions, average clusters, maximum clusters, concatenations, dropouts, and fully connected layers. Batch normalization is used extensively throughout the applied mode to trigger inputs.

• RESNET 50: One of the problems that occur with very deep networks is the well-known vanishing gradient. This problem occurs when the loss function gradients tend to zero after numerous partial derivatives in each training iteration are found. This process prevents the weights from being upgraded, thus, stopping the learning process. A Deep Residual Network (ResNet) uses a two-layer convolutional building block with an identity connection jumping over them [45]. It is similar to networks with convolution, pooling, activation, and fully connected layers stacked on top of each other. What makes it a residual network is the identity connection between the layers. In this way, the input of the model’s first layer then becomes the output of the last layer, the network must be able to predict any function it has previously learned with the input added to it. Gradients can flow directly through jump-back connections from the later layers to initial filters [45]. A specific ResNet network is instantiated by stacking these building blocks at the desired depth. ResNet-50 is a convolutional neural network that has 50 layers. The difference in accuracy between ResNet-18 and ResNet-50 was notable in a performance comparison [46]. For ResNet networks with depths of greater than 50, the increase in performance was moderate.

3.3. Training Strategy of CNN Models

In this paper, we used transfer learning and fine-tuning approaches (see Figure 5). Transfer learning is characterized by using a trained model in one domain as a starting point for another domain’s training phase, reducing the amount of data required for achieving acceptable results, consequently reducing the processing time [47]. The trained model provides a structure that previously acquired knowledge and can be used to solve new problems, in contrast to the approaches of classical machine learning algorithms [48,49].

The idea behind transfer learning is that if the model is trained on a large enough dataset, then it is likely that general characteristics common to visual systems could be learned. Therefore, it is possible to take advantage of these generated attribute maps to train models without constructing a model from scratch.

In this study, we investigated the performance of pre-trained VGG16, INCEPTION V3, and RESNET 50 models. All models are trained on ImageNet, a large dataset with more than 14 million annotated images and 1000 classes. These CNN architectures are widely addressed in several works involving gastric tissue imaging [10].

DeepHP was randomly divided into the training, validation, and test set. The training dataset was used for CNN training and for the adjustment of hyper-parameters. The validation set was employed to evaluate the CNN models during training and to make adjustments to the model. The test set was used to analyze the model’s performance through evaluation metrics (see Section 3.4) and verify the generalization of the network.
Figure 5. Overview of transfer learning approach. The frozen convolutional layers are highlighted in yellow, and the trainable layers of the models are highlighted in blue.

For feature extraction, we added a fully-connected classifier at the top. The pre-trained model is “frozen”, and only the classifier weights are updated during training. In this case, the convolutional base extracted all the resources associated with each image. We only train a classifier that determines the image class with that set of extracted features.

We used a grid search approach on the classifier parameters and evaluated it using 5-fold cross-validation for 20 epochs on a sample with 10% of the dataset. The parameters used in the grid search were the number of fully connected neurons, activation functions, number of layers, dropout, optimizer, and learning rate. For each epoch, 32 samples are randomly selected by using ADA and SGD with 0.01, 0.001, and 0.0001 learning rates to update the network weights. One epoch is over until the whole dataset is used. The sigmoid function was adopted to classify samples, and all networks added one or two full connect layers before the sigmoid layer. All training data were augmented by performing rotation transformation, zooming, horizontal flip, vertical flip, and scale normalization.

After evaluating the training results, we discard the less accurate models and re-apply the above methodology to the total dataset. We select the best combination for each model and then fine-tune it. For fine-tuning, we unlocked part of the convolutional layers. Thus, their weights are adjusted so that the model learns high-level features specific to the dataset. The parameters in the convolutional layers do not need to be changed. However, it is necessary to define how many convolutional layers will be unfrozen. We include an accuracy checkpoint as a monitor to save the most accurate weight set.

We develop our applications in Python, with the Application Programming Interface (API) Keras, working with TensorFlow. Keras is primarily focused on ease of use, modulation, and extension and has well-known models of famous high-performance networks, so it is a powerful tool for transfer learning tasks. We train our networks on a computer with an Intel Core i9 3.7/5.3 GHz processor, 16 GB RAM, 2 TB SATA HDD, 250 GB SSD, and NVIDIA RTX 3070 8 GB. For the total data, we use a computer with an AMD Threadripper 3960X 3.8/4.5 GHz processor, 128 GB RAM—8 × 16 GB, 10 TB SATA HDD, 250 GB SSD, NVIDIA RTX 3090 24 GB.
3.4. Evaluation Methods for CNN Models

For CNN model evaluation, we measured the Receiver Operating Characteristic Curve (ROC curve) and the Area Under Curve (AUC), the F1-score defined in Equation (7), and the Matthews correlation coefficient defined in Equation (10).

Based on the ground truth, the F1-score can be calculated as follows:

\[ F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \]  

where precision can be calculated by:

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

and recall by:

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

The Mathews correlation coefficient was also used as an evaluation metric. The MCC was first introduced by B.W. Mathews to predict the secondary protein structure [50] and can be calculated using the equation provided below:

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]  

4. Conclusions

Precision medicine is attaining attention. Deep learning-based applications are providing advancements in patient-level diagnostic support. Some diseases in which the diagnosis depends on a subjectivity analysis, such as histopathological exams, benefit from decision support systems using deep learning approaches. However, there are some challenges to fully implementing reliable decision support systems in this application domain. First, deep learning requires a large volume of data to achieve reliable results. Second, and no less critical, several works that use deep learning in the medical domain are not fully described or provide the datasets, impairing the experiments’ transparency and reproducibility.

Aiming to fill these gaps, we present DeepHP in this work, which is a curated public collection of histopathological images for detecting gastric mucosa with Helicobacter pylori infection. Moreover, we tested three deep learning architectures, namely VGG16, InceptionV3, and ResNet50. We conducted extensive computational experiments to test a transfer learning strategies.

Our results conclude that CNN models can detect HP infections and inflammatory profiles in gastric biopsies, providing adequate support for clinical diagnosis, regardless of whether physicians or specialists use the system. The use of accurate CNN models can help reduce the workload and highly laborious tasks of pathologists. In the future, we plan to use image collections with different staining methods and set a cutoff for classifying entire images rather than just patches and advancing parameter adjustment and data augmentation to improve these CNN architectures.

The future research possibilities are broad. Our dataset can be expanded and used by other research groups, considering that it is publicly available. The trained models can be used as a basis for other medical applications. Other CNN models can be explored. Finally, explainable AI methods can be investigated to attend to the FDA and other regulatory agencies’ requirements.

Author Contributions: Conceptualization, F.M.F.L., Á.R.-d.-S. and G.S.d.A.; Data curation, S.D. and G.S.d.A.; Formal analysis, W.G.e.G.; Funding acquisition, Á.R.-d.-S.; Methodology, W.G.e.G., M.H.P.d.S., L.M.B., F.M.F.L., S.D., Á.R.-d.-S. and G.S.d.A.; Software, W.G.e.G., M.H.P.d.S., L.M.B. and H.G.A.P.; Writing—original draft, W.G.e.G., and G.S.d.A.; Writing—review & editing, F.M.F.L., Á.R.-d.-S. and G.S.d.A. All authors have read and agreed to the published version of the manuscript.
**Funding:** This research was funded by PROPESP/UFPA for financial support and scholarships, and Fundação Amazônia Paraense de Amparo à Pesquisa—FAPESP (No. 008/2017 and No. BJT-2021/658671).

**Acknowledgments:** We would like to thank reviewers from Rede de Pesquisa em Genômica Populacional Humana/Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES (Bio. Computacional, No. 3381/2013); Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 2021, 71, 209–249. [CrossRef]

2. Thrift, A.P.; El-Serag, H.B. Burden of gastric cancer. *Clin. Gastroenterol. Hepatol.* 2020, 18, 534–542. [CrossRef] [PubMed]

3. Kusters, J.G.; Van Vliet, A.H.; Kuipers, E.J. Pathogenesis of Helicobacter pylori infection. *Clin. Microbiol. Rev.* 2006, 19, 449–490. [CrossRef] [PubMed]

4. Zamani, M.; Ebrahimtabar, F.; Zamani, V.; Miller, W.; Alizadeh-Navaei, R.; Shokri-Shirvani, J.; Derakhshan, M. Systematic review with meta-analysis: The worldwide prevalence of Helicobacter pylori infection. *Aliment. Pharmacol. Ther.* 2018, 47, 868–876. [CrossRef] [PubMed]

5. Wang, F.; Meng, W.; Wang, B.; Qiao, L. Helicobacter pylori-induced gastric inflammation and gastric cancer. *Cancer Lett.* 2014, 345, 196–202. [CrossRef]

6. Eusebi, L.H.; Telese, A.; Marasco, G.; Bazzoli, F.; Zagari, R.M. Gastric cancer prevention strategies: A global perspective. *J. Gastroenterol. Hepatol.* 2020, 35, 1495–1502. [CrossRef]

7. Coelho, L.G.V.; Marinho, J.R.; Genta, R.; Ribeiro, L.T.; Passos, M.d.C.F.; Zaterka, S.; Assumpção, P.P.; Barbosa, A.J.A.; Braga, L.L.; et al. IVth brazilian consensus conference on Helicobacter pylori infection. *Arq. Gastroenterol.* 2018, 55, 97–121. [CrossRef]

8. Garza-González, E.; Perez-Perez, G.I.; Maldonado-Garza, H.J.; Bosques-Padilla, F.J. A review of Helicobacter pylori diagnosis, treatment, and methods to detect eradication. *World J. Gastroenterol.* WJG 2014, 20, 1438. [CrossRef] [PubMed]

9. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal.* 2020, 63, 101693. [CrossRef] [PubMed]

10. Guimarães, P.; Keller, A.; Fehlmann, T.; Lammert, F.; Casper, M. Deep-learning based detection of gastric precancerous conditions. *Nutrocomp* 2020, 276, 2–22. [CrossRef]

11. Gonçalves, W.G.; dos Santos, M.H.d.P.; Lobato, F.M.F.; Ribeiro-dos Santos, Â.; de Araújo, G.S. Deep learning in gastric tissue diseases: A systematic review. *BMJ Open Gastroenterol.* 2020, 7, e000371. [CrossRef]

12. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *J. Imaging* 2020, 6, 52. [CrossRef]

13. Li, Z.; Togo, R.; Ogawa, T.; Haseyama, M. Semi-supervised learning based on tri-training for gastritis classification using gastric X-ray images. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5. [CrossRef]

14. Guimarães, P.; Keller, A.; Fehlmann, T.; Lammert, F.; Casper, M. Deep-learning based detection of gastric precancerous conditions. *Gut* 2020, 69, 4–6. [CrossRef]

15. Togo, R.; Yamamichi, N.; Mabe, K.; Takahashi, Y.; Takeuchi, C.; Kato, M.; Sakamoto, N.; Ishihara, K.; Ogawa, T.; Haseyama, M. Detection of gastritis by a deep convolutional neural network from double-contrast upper gastrointestinal barium X-ray radiography. *J. Gastroenterol.* 2019, 54, 321–329. [CrossRef] [PubMed]

16. Churchill, S.; Nomura, S.; Aoyama, K.; Nishikawa, Y.; Miura, M.; Shinagawa, T.; Takiyama, H.; Tanimoto, T.; Ishihara, S.; Matsuo, K.; et al. Application of convolutional neural networks in the diagnosis of Helicobacter pylori infection based on endoscopic images. *EBioMedicine* 2017, 25, 106–111. [CrossRef]

17. Nakashima, H.; Kawahira, H.; Kawachi, H.; Sakai, N. Artificial intelligence diagnosis of Helicobacter pylori infection using blue laser imaging-bright and linked color imaging: A single-center prospective study. *Ann. Gastroenterol.* 2018, 31, 462. [CrossRef]

18. Itoh, T.; Kawahira, H.; Nakashima, H.; Yata, N. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy images. *Endosc. Int. Open* 2018, 6, E139–E144. [CrossRef]

19. Shichijo, S.; Endo, Y.; Aoyama, K.; Takeuchi, Y.; Ozawa, T.; Takiyama, H.; Matsuo, K.; Fujishiro, M.; Ishihara, S.; Ishihara, R.; et al. Application of convolutional neural networks for evaluating Helicobacter pylori infection status on the basis of endoscopic images. *Scand. J. Gastroenterol.* 2019, 54, 158–163. [CrossRef] [PubMed]

20. Zheng, W.; Zhang, X.; Kim, J.J.; Zhu, X.; Ye, G.; Ye, B.; Wang, J.; Luo, S.; Li, J.; Yu, T.; et al. High Accuracy of Convolutional Neural Network for Evaluation of Helicobacter pylori Infection Based on Endoscopic Images: Preliminary Experience. *Clin. Transl. Gastroenterol.* 2019, 10, e00109. [CrossRef] [PubMed]
21. Lin, Y.W.; Lin, G.S.; Chai, S.K.D. Helicobacter Pylori Classification based on Deep Neural Network. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
22. Nakashima, H.; Kawahara, H.; Kawachi, H.; Sakani, N. Endoscopic three-categorical diagnosis of Helicobacter pylori infection using linked color imaging and deep learning: A single-center prospective study (with video). *Gastric Cancer*. 2020, 23, 1033–1040. [CrossRef]
23. Martin, D.R.; Hanson, J.A.; Gullapalli, R.R.; Schultz, F.A.; Sethi, A.; Clark, D.P. A deep learning convolutional neural network can recognize common patterns in injury pathology. *Arch. Pathol. Lab. Med.* 2020, 144, 370–378. [CrossRef]
24. Yu, H.; Yang, L.T.; Zhang, Q.; Armstrong, D.; Deen, M.J. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* 2021, 444, 92–110. [CrossRef]
25. Chopygina, V.; de Bruijne, M.; Pluim, J.P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* 2019, 54, 280–296. [CrossRef] [PubMed]
26. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* 2020, 21, 1–13. [CrossRef]
27. Wardhani, N.W.S.; Rochayani, M.Y.; Iriany, A.; Sulistyono, A.D.; Lestantyo, P. Cross-validation metrics for evaluating classification performance on imbalanced data. In Proceedings of the 2019 International Conference on Computer, Control, Informatics and Its Applications (ic3ina), Tangerang, Indonesia, 23–24 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 14–18.
28. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
29. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 2009, 21, 1263–1284.
30. Pearson, R.; Goney, G.; Shwaber, J. Imbalanced clustering for microarray time-series. In Proceedings of the ICML, Washington, DC, USA, 21–24 August 2003; Volume 3.
31. Harrison, C. FDA backs clinician-free AI imaging diagnostic tools. *Nat. Biotechnol.* 2018, 36, 673. [CrossRef] [PubMed]
32. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kattoo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 2018, 29, 1836–1842. [CrossRef]
33. Food and Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*; FDA: Silver Spring, MD, USA, 2019.
34. Fleiss, J.L.; Levin, B.; Paik, M.C. Statistical Methods for Rates and Proportions; John Wiley & Sons: Hoboken, NJ, USA, 2013.
35. Congalton, R.G.; Green, K. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices; CRC Press: Boca Raton, FL, USA, 2019.
36. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* 1977, 33, 159–174. [CrossRef]
37. Matsugu, M.; Morii, K.; Mitari, Y.; Kaneda, Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw.* 2003, 16, 555–559. [CrossRef]
38. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444. [CrossRef] [PubMed]
39. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* 1986, 323, 533–536. [CrossRef]
40. Alom, M.Z.; Taha, T.M.; Alnaser, M.; Saad, H.; Hassan, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 2018, 29, 1836–1842. [CrossRef]
41. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* 2019, 8, 292. [CrossRef]
42. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
43. Rangarajan, A.K.; Purushothaman, R. Disease classification in eggplant using pre-trained VGG16 and MSVM. *Sci. Rep.* 2020, 10, 2322. [CrossRef]
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014, arXiv:1409.1556.
45. Ji, Q.; Huang, J.; Ren, S.; Sun, Y. Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. *Algorithms* 2019, 12, 51. [CrossRef]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Canziani, A.; Paszke, A.; Culurciello, E. An analysis of deep neural network models for practical applications. *arXiv* 2016, arXiv:1605.07676.
48. Agrusa, A.S.; Gharibans, A.A.; Allegra, A.A.; Kunkel, D.C.; Coleman, T.P. A deep convolutional neural network approach to classify normal and abnormal gastric slow wave initiation from the high resolution electrogastrogram. *IEEE Trans. Biomed. Eng.* 2019, 67, 854–867. [CrossRef]
49. Kumar, B.S.; Ravi, V. A survey of the applications of text mining in financial domain. *Knowl.-Based Syst.* 2016, 114, 128–147. [CrossRef]
50. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) Protein Struct.* 1975, 405, 442–451. [CrossRef]