diagnostics

MDPI

Article

Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning

Ming Ping Yong [1], Yan Chai Hum [1], Khin Wee Lai [2], Ying Loong Lee [1], Choon-Hian Goh [1], Wun-She Yap [1] and Yee Kai Tee [1,]*

[1] Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia; leeyingl@utar.edu.my (Y.L.L.)
[2] Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia
* Correspondence: teeyeekai@gmail.com or teeyk@utar.edu.my

Abstract: Gastric cancer is a leading cause of cancer-related deaths worldwide, underscoring the need for early detection to improve patient survival rates. The current clinical gold standard for detection is histopathological image analysis, but this process is manual, laborious, and time-consuming. As a result, there has been growing interest in developing computer-aided diagnosis to assist pathologists. Deep learning has shown promise in this regard, but each model can only extract a limited number of image features for classification. To overcome this limitation and improve classification performance, this study proposes ensemble models that combine the decisions of several deep learning models. To evaluate the effectiveness of the proposed models, we tested their performance on the publicly available gastric cancer dataset, Gastric Histopathology Sub-size Image Database. Our experimental results showed that the top 5 ensemble model achieved state-of-the-art detection accuracy in all sub-databases, with the highest detection accuracy of 99.20% in the 160 × 160 pixels sub-database. These results demonstrated that ensemble models could extract important features from smaller patch sizes and achieve promising performance. Overall, our proposed work could assist pathologists in detecting gastric cancer through histopathological image analysis and contribute to early gastric cancer detection to improve patient survival rates.

Keywords: histopathology; gastric cancer; deep learning; convolutional neural network; transfer learning; ensemble model

check for updates

Citation: Yong, M.P.; Hum, Y.C.; Lai, K.W.; Lee, Y.L.; Goh, C.-H.; Yap, W.-S.; Tee, Y.K. Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning. *Diagnostics* **2023**, *13*, 1793. https://doi.org/10.3390/diagnostics13101793

Academic Editors: Manuel Scimeca, Alessandro Mauriello, Rita Bonfiglio, Nicola Fusco and Yanwu Xu

Received: 30 March 2023
Revised: 8 May 2023
Accepted: 14 May 2023
Published: 18 May 2023

1. Introduction

Gastric cancer is one of the most common cancers and leading causes of cancer-related mortality [1]. Gastric cancer is considered a single heterogeneous disease with several histopathologic characteristics [2], where the gastric cancer presents distinct subtype with different histologic appearance, making the detection a non-trivial task. The clinical gold standard of gastric cancer detection is histopathology screening of a biopsy or surgical specimen using a microscope to identify the cancerous features [3]. This is done conventionally by pathologists by manually screening the tissue biopsies, first by using a low magnification factor to search for potential cancerous region(s) with naked eyes. Once a suspicious region is identified, the pathologists will switch to a high magnification factor to analyze the details of the region. During the diagnostic procedure, the pathologists assess the gigapixel-sized whole slide image (WSI) by traversing the WSI to find the small abnormal region of interest (ROI) as described above repeatedly, to make diagnostic decisions.

However, this conventional and manual visual analysis of tissue biopsies by pathologists is extremely laborious, time-consuming, and subjective, where the conclusion drawn by a pathologist can be different from another. The correct analysis of histopathology is highly dependent upon the expertise and experience of the pathologists. This makes the manual histopathological analysis prone to human errors such as misdetection and misdiagnosis, coupled with a shortage of pathologists, leading to long backlogs in the processing of patient cases and consequently increases the likelihood of delayed cancer detection.

Since most gastric cancers are adenocarcinomas, there are no apparent symptoms in the early stage or may present with non-specific symptoms such as gastric discomfort which are often mistaken as gastric ulcers and gastritis [4]; this causes a delay in the gastric cancer detection. Early detection of gastric cancer is the key factor to reduce mortality [5]. This can be observed in patients with an early gastric cancer diagnosis and detection; they have a survival rate of above 90% [6]. When detected in the late stage, the survival rate reduces substantially to below 30% [7,8].

The limitations of the manual diagnostic workflow lead to the development of computer-aided diagnosis (CAD) to assist pathologists by making the diagnosis more efficient and autonomous. CAD is gaining attention and becoming more accessible nowadays due to the advancement in digital pathology, resulting in slide scanning quality improvement and cost reduction in digital storage [9]. In addition, these systems not only reduce the time and cost of cancer diagnosis but also the inter-pathologist variability in diagnostic decisions [10].

For gastric cancer detection using histopathological images, various CAD techniques have been explored based on classification and segmentation models. Machine learning is the conventional CAD approach used to perform gastric cancer detection. In this approach, the used models extract handcrafted features such as color, texture, and shape features for the detection [11–13]. The common machine learning classifiers are support vector machine (SVM), random forest, and Adaboost [14–16].

Later, the deep learning approach is introduced to automate feature selection. Many works have reported deep convolutional neural networks (CNN) achieve promising performance in histopathological image classification and segmentation tasks in cancer [17–19], metastasis [20,21], and gene mutation [22,23] analysis; some even reported performance comparable to pathologists' assessment [9,24–28].

However, a good deep learning model requires huge datasets for training to obtain the optimal model parameters. This has particularly affected the use of deep learning in the histopathology image analysis domain due to the exhaustive manual data labeling process of the histopathology images. The data augmentation strategies such as affine transformation, color transformation, and noise addition are generally applied to expand the small histopathology gastric cancer dataset [11,29].

Another commonly used technique to overcome small dataset problems is the transfer learning method. While some researchers proposed their CNNs and trained them from scratch [30,31], others adopted transfer learning to fine-tune pre-trained networks for gastric cancer detection [5,30].

Each pre-trained network has its unique strengths and limitations in extracting features due to variances in their architectures. Although the networks make correct predictions, each pre-trained network may have a different interpretation of an image, and different important features are extracted in the classification process. Utilizing any one of these networks results in the loss of important features that may be extracted by others; this may lead to insufficient important features or information to generate the correct predictions and hence lead to a lower classification performance.

To address the shortcoming in the feature extraction process by the individual pre-trained networks, this work proposed ensemble models for histopathology gastric cancer detection. The ensemble models exploit the strength and overcome the limitation of each pre-trained network by combining decisions of multiple pre-trained networks in the classification process; therefore, they can generate a larger amount of important features or information that are required to make the correct predictions and subsequently improve the classification performance. This is especially useful for the histopathology domains because the WSIs are generally in high resolution; processing them directly would require high computational power; thus, they are commonly cropped into smaller patches with lower resolution for CAD. Although the smaller patches are more computationally friendly, it comes at a cost of lower performance due to less useful features or information in the images. The proposed deep ensemble learning has great potential to overcome this pressing issue in histopathological image analysis.

To verify the effectiveness of our proposed models, their performance was evaluated on the new publicly available gastric cancer dataset Gastric Histopathology Sub-size Image Database (GasHisSDB) [32].

The main contributions of this paper are that (1) deep ensemble learning models that are effective for gastric cancer detection are developed. This is proven by the better performance compared to the state-of-the-art studies on the GasHisSDB dataset; (2) deep ensemble learning can still accurately classify the gastric histopathological images with lower resolution. Therefore, it is feasible to reduce the specifications of the digital scanner, data storage, and high computational server required in the histopathology tasks, potentially translating to higher likelihood of early gastric cancer detection to improve patient survival rate.

2. Related Works

The classical machine learning approach based on handcrafted feature extractions was used in automating histopathology tasks initially. Doyle et al. [33] extracted various combinations of handcrafted textural and graph features such as gray level features, Haralick features, Gabor filter features, the Voronoi diagram, Delaunay triangulation, minimum spanning tree, and nuclear features. After that, the authors applied spectral clustering algorithms as dimensionality reduction methods to filter the useful features before passing them to SVM to classify whether the images are normal or breast cancer. The model achieved an accuracy of 95.8% in cancerous image detection and 93.3% in cancer image grading. In the work of Kather et al. [34], six distinct sets of handcrafted texture descriptors including lower-order and higher-order histogram features, local binary patterns, gray-level co-occurrence matrix, Gabor filters, and perception-like features were combined into a feature set; after that, various classifiers including the 1-nearest neighbor, linear SVM, radial-basis function SVM, and decision trees were used for the colorectal image binary and multiclass classification. The proposed work managed to achieve 98.6% accuracy in the binary classification and 87.4% accuracy in the multiclass study. Although the classical machine learning approach can achieve promising performance, it requires in depth expertise in the histopathology domain to design meaningful features, which serve as its shortcoming and barrier to developing an effective machine learning model.

To address this problem, deep learning approach is introduced for histopathology task automation. Unlike machine learning, deep learning models do not require handcrafted features as the input; they can learn the required features automatically. However, a huge dataset is usually needed for the deep learning models to learn the features effectively and then achieve a high performance.

Data augmentation and transfer learning are two common methods used to address the huge dataset requirement in training deep learning models. The previous generates artificial samples to expand the dataset. In the work of Sharma and Mehra [35], the dataset was augmented using flipping, translation, scaling, and rotation technique; Han et al. [36] balanced the dataset using the augmentation methods including intensity change, rotation, and flipping; Joseph et al. [37] applied translation, scaling, flipping, and rotation with constant fill mode to expand the dataset. The model accuracies improved by 2.76–12.28% across various magnifications in [35], 3.4% at the image level and 5.8% at the patient level in [36], and 4.52–8.17% across various magnifications in [37] in the respective tasks after the data augmentation.

The second method to overcome the huge dataset requirement is transfer learning, where a model that has been trained for one task is applied as a starting point of a model to perform a different task. In the work of Al-Haija et al. [38], the pre-trained ResNet50 was fine-tuned for the breast cancer classification task; Mehra [39] compared the transfer learning and training from scratch methods using three models which are VGG-16, VGG-19, and ResNet-50; Celik et al. [40] proposed transfer learning using the pre-trained networks DenseNet-161 and ResNet-50. The pre-trained networks accuracies improved by 5.9–14.76% in [38], 12.67% (between best performing models) in [39], and 1.96–6.73% in [40] in the respective tasks over the custom CNNs or training the models from scratch.

Although the methods above have achieved relatively good performance in the histopathological image analysis, there is another notable method called ensemble learning that can be integrated with these methods to further improve the classification performance. Ensemble learning involves aggregating the output decisions of multiple base models, which would be the pre-trained networks in this case, through relatively simple ensemble strategies to make the final predictions. The intuition behind the ensemble model is that each base model may have its limitation in feature extraction despite its good performance, and these limitations can be overcome through the strength of the other base models. Hence, by combining multiple base models, the ensemble model has a wider coverage of extracted features, resulting in better performance.

For instance, Ghosh et al. [41] proposed an ensemble model concatenating the results of DenseNet-121, InceptionResNetV2, Xception, and custom CNN to classify 112,180 colorectal images, which are resized into 100 × 100 pixels, into multiple classes. Different weights were assigned to the results of each base model depending on their individual performance. The ensemble model ultimately achieved 99.13% balanced accuracy. In the work of Zheng et al. [42], the weighted voting strategy was used as ensemble method to aggregate pre-trained networks including VGG-16, Xception, ResNet-50, and DenseNet-201 in performing breast cancer multiclass classification on 7909 images across four magnifications, achieving accuracy 98.90%. Paladini et al. [43] proposed using the feature concatenation strategy to aggregate the feature outputs of pre-trained networks including ResNet-101, ResNeXt-50, Inception-V3, and DenseNet-161 and consequently processed the aggregated feature vectors through fully connected and classification layers for the colorectal image classification using the dataset consists of 150 × 150 pixels images, achieving an accuracy of 96.16%. The ensemble models accuracies improved by 1.83–2.16% in [41], 0.1–5.25% in [42], and 0.74–2.18% in [43] in the respective tasks over their corresponding base models.

A WSI can be as large as 100,000 × 100,000 pixels; it is costly and time-consuming to annotate the WSI in detail. A common method to process the WSI is to crop it into smaller patches for artificial intelligence training and classification. Downsizing the WSI prior to cropping it into smaller patches is usually conducted for resource constraint centers. This comes at the cost of lower classification performance because the smaller patch size contains less information for classification purposes. Therefore, the selection of patch size demands the consideration of trade-off between computational power and classification performance.

With the promising performance shown by the ensemble models supported by its capability of extracting many important features from multiple base models, the ensemble models have the potential to extract sufficient important features from the smaller patch size yet achieve promising performance. This can have significant impact in making WSI with lower resolution to be more accessible to correct classification by deep learning models, consequently reducing the specification of the digital scanner, data storage, and high computational server required in the histopathology tasks. This would translate to more efficient and autonomous histopathological diagnosis, leading to lower likelihood of delayed cancer detection.

In this study, the GasHisSDB gastric dataset which consists of three sub-databases of patches, being 80 × 80 pixels, 120 × 120 pixels and 160 × 160 pixels, would be used to compare the performance of our proposed ensemble models with the state-of-the-art studies.

3. Materials and Methods

3.1. Dataset

The GasHisSDB dataset was prepared through a collaboration between Longhua Hospital Shanghai University of Traditional Chinese Medicine, Northeastern University, and

Liaoning Cancer Hospital and Institute [32]. The dataset consists of a total of 245,196 patches derived from 600 WSI of 2048 × 2048 pixels. The patches are divided into two classes: normal and abnormal class. The dataset distribution includes 97,076 abnormal image patches and 148,120 normal image patches. The patches were cropped from the WSI at three patch sizes: 80 × 80 pixels, 120 × 120 pixels, and 160 × 160 pixels. These patches of three different sizes were divided into three sub-databases.

The paper applied different algorithms to obtain patches from normal and abnormal tissue regions in WSI, respectively. For the normal tissue region, the patches were directly cropped from the normal pathological section whereas the cancerous tissue region was selected from the abnormal pathological section. Each cropped ROI would have a labeled ground truth. Cancerous image patches with less than 50% cancerous region according to the ground truth map were removed. The WSIs are hematoxylin and eosin (H&E) stained and have a magnification of ×20. The patches were rotated randomly, and dataset orders were scrambled to reduce the correlation between the patches that originated from a similar WSI. The dataset summary and samples are illustrated in Figure 1.

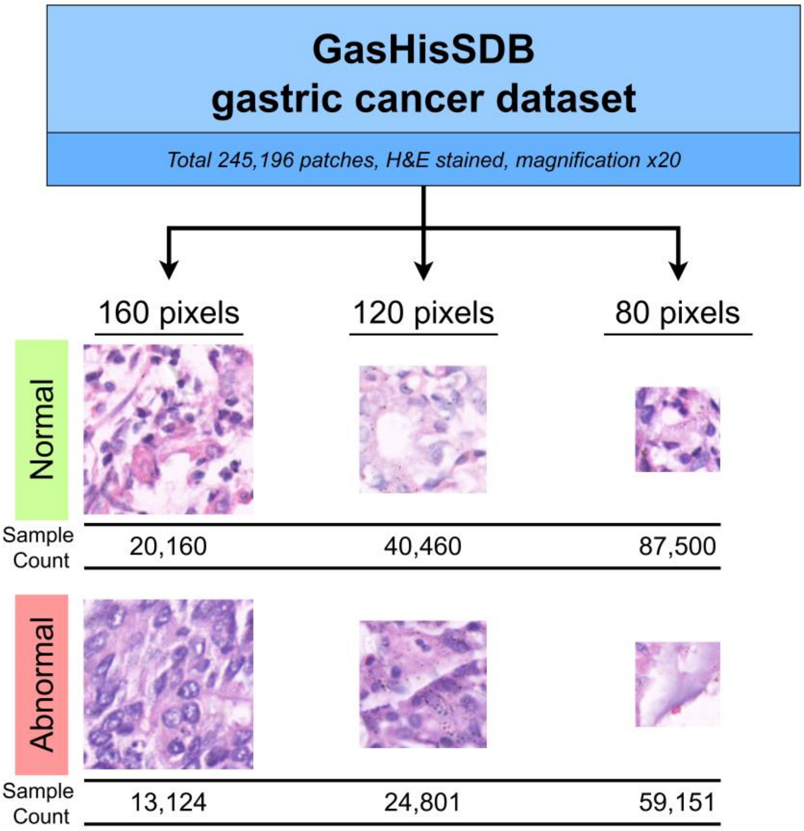

Figure 1. GasHisSDB dataset samples and summary.

3.2. Methodology Overview

This paper proposed CNN architectures based on transfer learning and ensemble models to perform patch binary classification tasks for gastric cancer detection. The workflow is decomposed into 4 stages: (i) preprocessing dataset with empty patch removal and data augmentation, (ii) fine-tuning pre-trained networks or base models, (iii) selecting the best performing base models to combine into ensemble models, and (iv) evaluating and visualizing proposed models using various metrics and class activation map.

3.3. Dataset Pre-Processing

Dataset pre-processing can generate a more distributed dataset to assist the subsequent model training for better model performance. The original dataset contains many empty patches that will deteriorate the model's performance. Therefore, the dataset was first preprocessed by removing empty patches, followed by data augmentation to expand the available data for model training.

3.3.1. Empty Patch Removal Process

We removed the empty patches from the dataset as these patches are non-informative. The empty patches were defined as patches that contained more than 10% pixels with RGB intensity value more than 230 across all channels. The percentage of removed patches are approximately 4.91% in the 80-pixel sub-database, 4.01% in the 120-pixel sub-database, and 2.86% in the 160-pixel sub-database. The number of samples before and after empty patch removal, respectively, are 146,615 and 139,415 samples in the 80-pixel sub-database, 65,261 and 62,645 samples in the 120-pixel sub-database, and 33,284 and 32,333 samples in the 160-pixel sub-database. After the empty patches were discarded, the rest of the patches were sent for data augmentation.

3.3.2. Data Augmentation

In this process, affine transformation including rotation at an interval of 90°, and horizontal or vertical flipping to the training images was applied to expand the dataset. The generated augmented training set became twice the size of the original training set. The number of samples before and after augmentation, respectively, are 55,766 and 111,532 samples in the 80-pixel sub-database, 25,058 and 50,116 samples in the 120-pixel sub-database, and 12,933 and 25,866 samples in the 160-pixel sub-database.

3.4. Pre-Trained Networks as the Base Models

We evaluated several commonly used pre-trained networks, including InceptionV3, Xception, MobileNet, MobileNetV2, DenseNet121, DenseNet169, EfficientNetB0, and EfficientNetB1, to identify and select the best-performing pre-trained networks as base models for the ensemble models in subsequent sections.

3.4.1. InceptionV3

InceptionV3 [44] was a CNN proposed by Szegedy et al. in 2015. This deals with the problem of conventionally increasing model size, leading to too many parameters and computational inefficiency by using factorized convolution. InceptionV3 computes multiple convolutions of different kernel sizes parallelly in a block, contributing to the strength of recognizing features at different scales. InceptionV3 also excels at overcoming vanishing gradient problems using auxiliary classifiers.

3.4.2. Xception

Xception [45] was a model proposed by Chollet. F. in 2017. This model is inspired by Inception, but it outperformed InceptionV3 on the ImageNet and JFT datasets when using the same number of parameters. Xception adds depthwise separable convolutions and residual connection into the model, with depthwise separable convolutions replacing the Inception modules. The depthwise separable convolutions used in Xception are modified, with pointwise convolution followed by depthwise convolution.

3.4.3. DenseNet Family

DenseNet [46] was a feedforward CNN proposed by Huang et al. in 2017. The model connects feature maps from all preceding layers as the inputs to all subsequent layers in each dense block. The purposes of this architecture are to reduce the vanishing gradient issue, promote feature reuse, and reduce the number of parameters. The feature maps from different layers in a dense block are combined through concatenation, and the downsampling process that cannot be carried out in a dense block for feature concatenation purposes using convolution and pooling operations.

3.4.4. EfficientNet Family

EfficientNet [47] was a model proposed by Tan, M. and Le, Q. in 2019. The main idea is under the resource constraint situation, scaling up models to achieve better accuracy is not feasible; the alternative of balancing the network depth, width, and resolution in the model can be used to improve the performance of the models. The work proposed using the compound scaling method to maintain efficiency in order to achieve better accuracy. This method scales up the baseline model, EfficientNetB0, uniformly across all dimensions to target resource constraints, generating a family of models from EfficientNetB0 to EfficientNetB7.

3.4.5. MobileNet Family

MobileNet [48] was a CNN proposed by Howard et al. in 2017. It is suitable for mobile and embedded vision applications. This model is lightweight and efficient because it has a smaller number of parameters compared to regular CNNs using the depthwise separable convolutions. MobileNetV2 [49] improves upon MobileNet by introducing the inverted residual blocks and bottlenecking features, leading to a significantly lower number of parameters and faster processing time.

3.5. Transfer Learning

Transfer learning using the CNN models mentioned in the previous section were applied. First, their weights were initialized to the weights pre-trained on the ImageNet dataset; then, the models were fine-tuned using the augmented training set.

3.6. Ensemble Models Architecture

Three ensemble strategies which are majority voting, unweighted averaging, and weighted averaging were used to build the ensemble models, and their performance was compared.

For each sub-database, the best-performing base models according to their validation set accuracies were selected to combine into ensemble models. The number of base models was chosen to be either three or five.

The majority voting strategy aggregates the prediction classes from base models through the voting process. The class with the highest votes was selected as the ensemble prediction class. The unweighted averaging strategy aggregates the prediction class probabilities from base models by averaging the base model prediction class probabilities for all classes. The class with the highest ensemble prediction probability was selected as the ensemble prediction class. For the weighted averaging method, unlike the unweighted averaging method, the ensemble prediction probabilities for all classes were averaged using an assigned specific weight for each base model under each class. The weights were optimized using feedforward neural networks in contrast to tuning them manually as hyperparameters.

3.7. Experiment Setting

The GasHisSDB dataset was split into train, validation, and test datasets with a train-validation-test ratio of 4:2:4 after the empty patch removal process. The stratified splitting was applied so that the ratio between the two image classes is balanced in all the datasets. The dataset distribution is shown in Table 1.

The pre-trained networks or base models employed were originally trained to classify 1000 classes of images on the ImageNet dataset. To adapt the models for our binary classification task, the original output softmax classification layers which have 1000 nodes were replaced with 2 nodes, corresponding to the binary classes. The remaining parts of the models were not modified.

Table 1. GasHisSDB dataset distribution after data pre-processing (empty patch removal and data augmentation).

Database	Number of Samples		
	Augmented Training Set	Validation Set	Testing Set
80-pixels	111,532	27,883	55,766
120-pixels	50,116	12,529	25,058
160-pixels	25,866	6466	12,934

For training parameters, each network was trained for 30 epochs, and the batch size was set to 20. For the backpropagation setting, the stochastic gradient descent (SGD) optimizer, categorical cross-entropy loss, and default learning rate of 0.01 were used. All layers in the networks including pre-trained layers were unfrozen and set to be trainable. The weight of the models at the epoch with the highest validation accuracy was chosen as the final representation of these models.

To implement the weighted averaging ensemble strategy in the ensemble model section, feedforward neural networks were used. The ensemble probabilities for the binary classes were obtained using two feedforward neural networks. Each network deals with a specific class by processing probabilities of this class from the multiple base models. The network consists of input and output layers only with no hidden layers, and the sigmoid activation function was used. Base model probabilities were fed as the inputs, and ensemble probability was used as the output in each network. The ensemble class probabilities from the two feedforward neural networks were aggregated. The class with the higher ensemble probability was selected as the ensemble prediction class. For model initialization, the model weights were set to zero. The training parameters of these feedforward neural networks were 5 epochs using the Adam optimizer, batch size of 20, and a default learning rate of 0.01.

3.8. Model Evaluation and Visualization

3.8.1. Evaluation Metrics

The metrics used for performance evaluation are accuracy (1), precision (2), recall (3), specificity (4), F1-score (5), area under the curve (AUC), and categorical cross-entropy loss. The abnormal or cancerous patches are labeled as positive samples, and normal or healthy patches are labeled as negative samples. True positive, true negative, false positive, and false negative cases are annotated as *TP*, *TN*, *FP*, and *FN*, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

3.8.2. Prediction Visualization

Visualizing model decisions on the images can help us to understand the model behaviors better. In this work, the feature maps of the models were visualized using the Gradient-weighted Class Activation Mapping (Grad-CAM) technique. It is a technique used to produce a coarse localization map using gradients flowing into final convolutional layers in the model to highlight important regions in the images used by each base model to make a classification [50].

The whole processing framework of the proposed ensemble learning models is summarized in Figure 2.

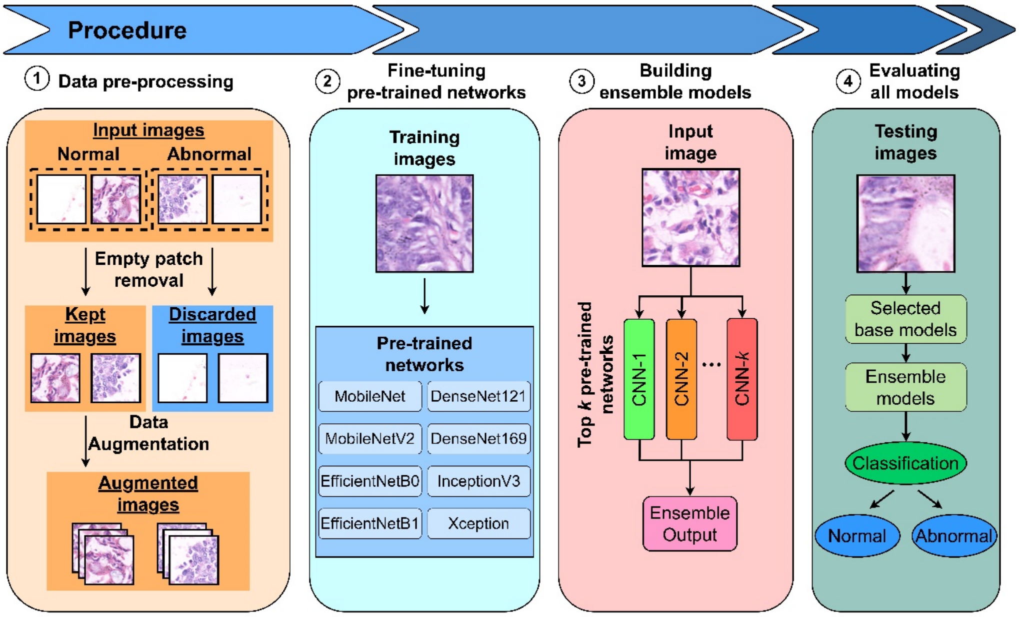

Figure 2. The architecture of the proposed ensemble models.

4. Results

The accuracy of the different fine-tuned base models on the validation set of the 80-, 120-, and 160-pixel categories is shown in Table 2. EfficientNetB1 had the highest accuracy in the 80-pixel sub-database category, achieving 96.75%; DenseNet169 achieved the highest accuracy (98.21%) in the 120-pixel sub-database, and DenseNet121 was the most accurate network (99.10%) in the 160-pixel sub-database. A more complete performance comparison of fine-tuning base models measured using other evaluation metrics is available in the Supplementary Materials (Tables S1–S3).

Table 2. Performance of base models on the 80-, 120-, and 160-pixel sub-database validation sets. The best-achieved results are bold.

Model	Accuracy (%)		
	80-Pixels	120-Pixels	160-Pixels
MobileNet	96.06	97.20	97.99
MobileNetV2	95.49	97.51	98.39
EfficientNetB0	**96.75**	97.72	98.48
EfficientNetB1	96.66	97.82	98.50
DenseNet121	96.65	98.12	**99.10**
DenseNet169	96.73	**98.21**	98.93
InceptionV3	94.75	96.72	98.24
Xception	95.80	97.14	97.79

The top 3 and 5 fine-tuned base models based on the validation accuracy are listed in Table 3. The top 3 models were the variants of EfficientNet and DenseNet. For the top 5 models, they were variants of EfficientNet, DenseNet, and MobileNet. These were selected as the base models of the deep ensemble learning for the testing set.

Table 3. Ranking of the top 3 and 5 fine-tune base models based on the validation accuracy; they are selected as the base models to be used in the ensemble models.

Ranking	80-Pixels	120-Pixels	160-Pixels
1	EfficientNetB0	DenseNet169	DenseNet121
2	DenseNet169	DenseNet121	DenseNet169
3	EfficientNetB1	EfficientNetB1	EfficientNetB1
4	DenseNet121	EfficientNetB0	EfficientNetB0
5	MobileNet	MobileNetV2	MobileNetV2

The performance of the different fine-tuned base models and ensemble models on the testing set of the 80-, 120-, and 160-pixel categories is shown in Tables 4–6. For the 80-pixel sub-database (Table 4), the best performing base model on the testing set was DenseNet169, achieving an accuracy of 96.67% followed by DenseNet121 and the EfficientNet variants. The ranking was slightly different compared to the performance on the validation set (Table 3), but the performance of the top base models was not too different (less than 1% difference in accuracy). When the top 3 and 5 base models were used for ensemble models according to the selections in Table 3, Ensemble-UA5 had the highest accuracy in the 80-pixel sub-database category, achieving an accuracy of 97.72% and the best AUC, precision, recall, specificity, and F1-score compared to all the tested base and ensemble models.

Table 4. Performance of the different deep learning models on the 80-pixel sub-database testing set. The best-achieved results are bold. For the ensemble learning models, WA stands for weighted averaging; UA stands for unweighted averaging, and MV stands for majority voting, and the 3 and 5 at the end of the ensemble models refer to top 3 or 5 base models. All metrics are measured in % unit.

Model	Accuracy	AUC	Precision	Recall	Specificity	F1-Score
MobileNet	95.82	95.73	94.90	95.15	96.30	95.02
MobileNetV2	95.29	94.87	96.36	92.26	97.48	94.27
EfficientNetB0	96.47	96.46	95.26	96.39	96.53	95.82
EfficientNetB1	96.50	96.41	95.83	95.83	96.99	95.83
DenseNet121	96.61	96.30	97.42	94.41	98.20	95.89
DenseNet169	**96.67**	**96.70**	**95.26**	**96.88**	**96.52**	**96.07**
InceptionV3	94.56	94.55	92.71	94.47	94.63	93.58
Xception	95.48	95.40	94.34	94.92	95.88	94.63
Ensemble-WA3	97.56	97.51	96.97	97.22	97.80	97.09
Ensemble-WA5	97.69	97.59	97.54	96.95	98.23	97.24
Ensemble-UA3	97.59	97.57	96.80	97.47	97.67	97.13
Ensemble-UA5	**97.72**	**97.65**	**97.39**	**97.18**	**98.12**	**97.28**
Ensemble-MV3	97.49	97.47	96.66	97.38	97.57	97.02
Ensemble-MV5	97.66	97.59	97.32	97.10	98.07	97.21

The performance of the different base models and the ensemble models on the 120-pixel sub-database testing set is presented in Table 5. DenseNet169 again was the best performing base model, achieving an accuracy of 98.17% followed by DenseNet121 and the EfficientNet variants. The ranking of the top 3 and 5 base models was the same in the testing (Table 5) and validation set (Table 3). When the top performing base models were used in the ensemble learning, Ensemble-WA5 achieved the highest accuracy (98.68%) and the best AUC, precision, recall, specificity, and F1-score compared to all the tested base and ensemble models in the 120-pixel sub-database.

Table 5. Performance of the different deep learning models on the 120-pixel sub-database testing set. The best-achieved results are bold. For the ensemble learning models, WA stands for weighted averaging; UA stands for unweighted averaging, and MV stands for majority voting, and the 3 and 5 at the end of the ensemble models refer to top 3 or 5 base models. All metrics are measured in % unit.

Model	Accuracy	AUC	Precision	Recall	Specificity	F1-Score
MobileNet	97.12	96.88	96.88	95.76	98.00	96.32
MobileNetV2	97.54	97.59	96.00	97.82	97.35	96.90
EfficientNetB0	97.66	97.66	96.42	97.68	97.64	97.04
EfficientNetB1	97.76	97.67	97.09	97.23	98.10	97.16
DenseNet121	97.87	97.70	97.72	96.86	98.53	97.29
DenseNet169	**98.17**	**98.02**	**98.04**	**97.30**	**98.74**	**97.67**
InceptionV3	96.63	96.45	95.80	95.63	97.27	95.71
Xception	97.03	96.86	96.43	96.03	97.69	96.23
Ensemble-WA3	98.52	98.36	98.59	97.63	99.09	98.11
Ensemble-WA5	**98.69**	**98.59**	**98.54**	**98.13**	**99.06**	**98.33**
Ensemble-UA3	98.53	98.42	98.36	97.90	98.94	98.13
Ensemble-UA5	98.68	98.61	98.38	98.27	98.95	98.32
Ensemble-MV3	98.47	98.35	98.32	97.78	98.91	98.05
Ensemble-MV5	98.64	98.57	98.32	98.23	98.91	98.27

Table 6. Performance of the different deep learning models on the 160-pixel sub-database testing set. The best-achieved results are bold. For the ensemble learning models, WA stands for weighted averaging; UA stands for unweighted averaging, and MV stands for majority voting, and the 3 and 5 at the end of the ensemble models refer to top 3 or 5 base models. All metrics are measured in % unit.

Model	Accuracy	AUC	Precision	Recall	Specificity	F1-Score
MobileNet	98.00	97.72	98.75	96.28	99.17	97.50
MobileNetV2	98.42	98.36	98.03	98.05	98.66	98.04
EfficientNetB0	98.33	98.29	97.83	98.05	98.52	97.94
EfficientNetB1	98.36	98.32	97.85	98.11	98.53	97.98
DenseNet121	**98.68**	**98.58**	**98.66**	**98.07**	**99.09**	**98.36**
DenseNet169	98.57	98.36	99.20	97.25	99.47	98.22
InceptionV3	97.85	97.83	97.02	97.69	97.96	97.36
Xception	97.43	97.34	96.74	96.91	97.78	96.82
Ensemble-WA3	98.94	98.78	99.44	97.94	99.62	98.68
Ensemble-WA5	99.16	99.09	99.19	98.72	99.45	98.96
Ensemble-UA3	99.03	98.93	99.13	98.45	99.42	98.79
Ensemble-UA5	**99.20**	**99.14**	**99.23**	**98.80**	**99.48**	**99.01**
Ensemble-MV3	98.97	98.88	99.06	98.40	99.36	98.73
Ensemble-MV5	99.13	99.07	99.10	98.76	99.39	98.93

In the 160-pixel sub-database testing set (Table 6), DenseNet121 was the best performance base model, obtaining an accuracy of 98.68% followed by the DenseNet169, MobileNetV2, and EfficientNet variants. The ranking of the top 3 and 5 base models was not the same in the testing (Table 6) and validation (Table 3) set, but their performance was close to each other (less than 1% difference in accuracy). When the top 3 and 5 base models were used for ensemble models according to the selections in Table 3, Ensemble-UA5 was the most accurate network, achieving an accuracy of 99.20% and the best AUC, precision, recall, specificity, and F1-score compared to all the tested base and ensemble models in the 160-pixel sub-database.

5. Discussion

5.1. Performance Analysis of the Base and Ensemble Models

From the performance of the various base and ensemble models in Tables 4–6, three interesting trends were observed. Firstly, the ensemble models always outperformed the fine-tuned base models under each sub-database. The lowest base model accuracy (94.56%) was recorded by InceptionV3 in the 80-pixel sub-database whereas the highest accuracy (98.68%) base model was DenseNet121 in the 160-pixel sub-database. Overall, all the base models were able to perform relatively well in gastric cancer detection.

The Grad-CAM maps highlighting the important regions in the histopathological images used by the top 5 base model to make a classification are plotted in Figure 3. It is interesting to notice that all the base models relied on different regions in the images to classify them. Therefore, when the ensemble models were applied, more features could be analyzed, leading to consistently better results when compared to the individual base model.

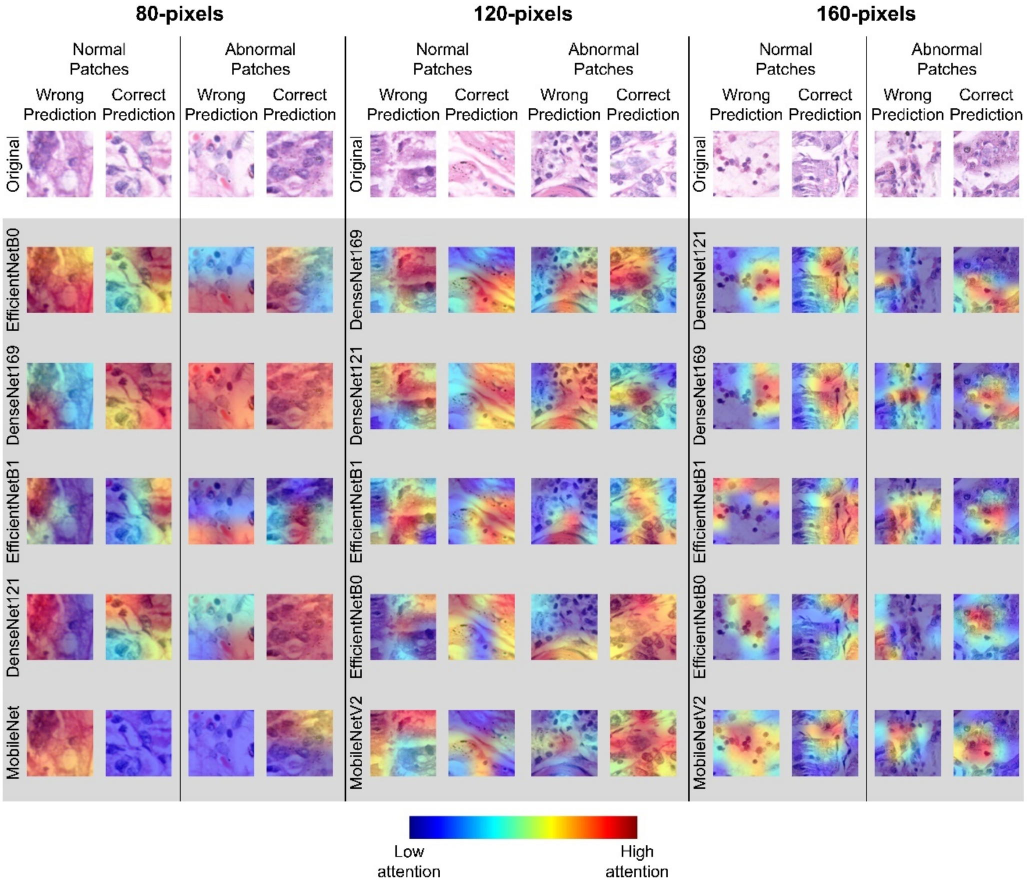

Figure 3. Samples and corresponding Grad-CAM heatmaps generated by the top five base models under all the three sub-databases.

The second interesting observation was that the top 5 ensemble model was consistently better than the top 3 ensemble model in each of the sub-databases tested when the same ensemble strategy was used. This was because all the base models were able to perform relatively good (>94.56%) and different base models focused on different important regions in the images for the classification. Thus, when more base models were included in the ensemble learning, higher accuracy was obtained.

The third interesting observation was the best performing model accuracy increased as the patch size increased as shown in Figure 4. The best accuracy of the base model increased from 96.67% in the 80-pixel sub-database to 98.68% in the 160-pixel sub-database. The same was also observed for the ensemble model, increasing from 97.72% to 99.20% when the number of pixels of the sub-databases changed from 80 to 160. This was not surprising given more useful features would be available in the images when the image resolution increased.

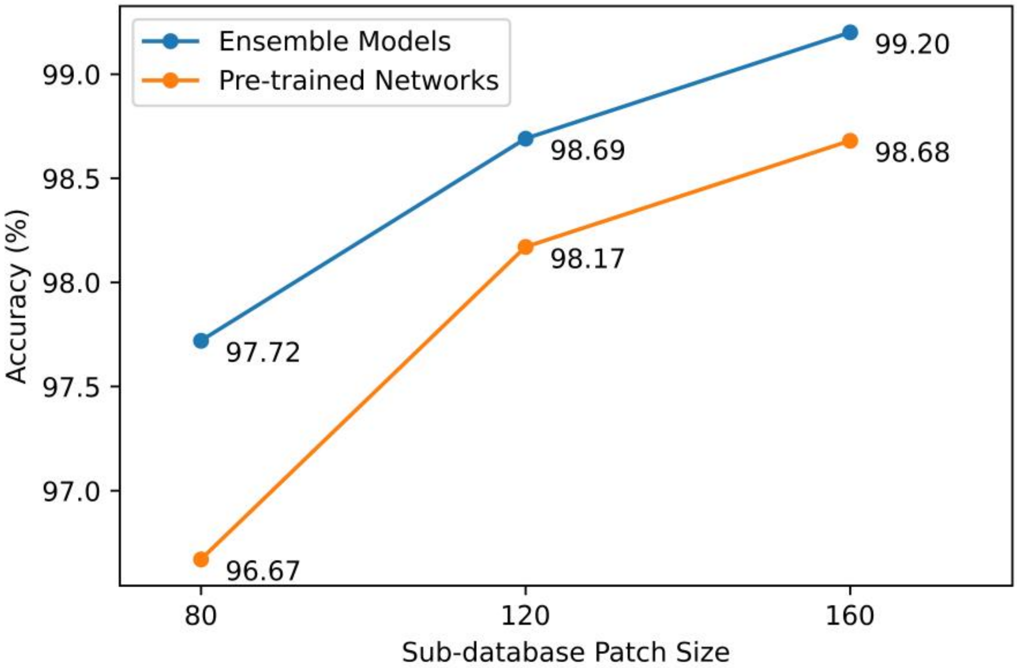

Figure 4. Testing accuracies of best-performing base models and ensemble models across all sub-databases.

5.2. Performance Analysis of the Proposed Models and the State-of-the-Art Studies on the GasHisSDB Dataset

Table 7 presents the details of the experiment and the best-performing models from our study and the previous state-of-the-art studies on the GasHisSDB dataset. From the table, all the studies used either 20% or 40% of the data as the testing set. We chose the latter to assess the performance of our proposed models.

Table 7. Performance comparison of our work and the previous state-of-the-art studies on the GasHisSDB dataset. The best-achieved results are bold. Only our best proposed ensemble models are included in this table, but all the proposed ensemble models achieve detection accuracy higher than the performance reported in the literature.

Paper	**Training /Validation/ Testing**	**Dataset Pre-Processing**	**Model Details**	**Accuracy (%)**		
				80-Pixels	**120-Pixels**	**160-Pixels**
[32]	40%/40%/20%	-	VGG16	96.12	96.47	95.90
			ResNet50	96.09	95.94	96.09
[51]	40%/20%/40%	-	InceptionV3 trained from scratch	-	-	98.83 ± 0.05
			InceptionV3 + ResNet50 (feature concatenation)	-	-	98.80 ± 0.12
[52]	60%/20%/20%	-	Local-global feature fuse network	-	-	96.81
[4]	80%/-/20%	-	MCLNet based on ShuffleNetV2	96.28	97.95	97.85
Our study (only the best model is listed)	40%/20%/40%	Data augmentation, empty patch removal	EfficientNetB0 + EfficientNetB1+ DenseNet121 + DenseNet169 + MobileNet (unweighted averaging)	**97.72**	98.68	**99.20**
			EfficientNetB0 + EfficientNetB1+ DenseNet121 + DenseNet169 + MobileNetV2 (weighted averaging)	97.69	**98.69**	99.16

All our ensemble models managed to obtain a gastric cancer detection accuracy higher than the reported performances in the literature (Tables 4–6), but only our best proposed ensemble model for each of the sub-database are listed in Table 7. The highest performance improvement of our proposed ensemble model when compared to the literature was in the 80-pixel sub-database, where our best obtained detection accuracy was 97.72, approximately 1.44% (97.72–96.28%) higher than the best reported accuracy in the literature; this was followed by 0.74% improvement in the 120-pixel sub-database and 0.37% in the 160-pixel sub-database. As discussed above, as the resolution of the images increased, more useful features would be available in the images; thus, the performance of the best model in the literature would be better, leading to a smaller percentage of improvement when compared to our proposed model.

More comprehensive performance comparison tables, which include performance metrics such as AUC, precision, recall, specificity, and specificity, are available in the Supplementary Materials (Tables S4–S6). These supplementary tables show that our models not only outperformed the previous state-of-the-art studies in terms of the accuracy but also in other metrics.

Our proposed ensemble models managed to obtain the state-of-the-art gastric cancer detection accuracy even with the ratio of training/validation/testing as 40%/20%/40%, meaning we used equal or lesser data for training and validation compared to the reported studies. This performance highlighted the superiority and robustness of our proposed models. The good performance of our work could be attributed to (1) the pre-processing steps taken and (2) the ensemble models used. In our work, the histopathological dataset was first preprocessed by removing empty patches to filter the non-informative images, thus improving the model feature learning. After that, data augmentation was applied to expand the available data for model training. These pre-processing steps were not taken by previous work in the literature.

As shown by the Grad-CAM maps in Figure 3, each base model focuses on different regions in the image to make a classification. Our proposed ensemble models allow the final decision to rely on wider image features thus resulting in the best gastric cancer detection accuracy on the GasHisSDB dataset.

5.3. Extended Experiments

Our proposed ensemble models were able to achieve the best performance in the binary classification on the GasHisSDB dataset. To prove the effectiveness and robustness of our proposed ensemble models and also to show that the proposed work is not sample/dataset limited, we further experimented these models on a different histopathology dataset named Histology Image Collection Library (HICL) histopathology larynx dataset [53,54]. This is a multi-class dataset which consists of Grade I, II, and III tumors and has a total of 224 images across all three classes. The immunohistochemistry (IHC) stains were applied on the images, different from the GasHisSDB dataset which used H&E stains.

Each image is presented in various resolutions, and we selected the 534 × 400 pixel and 1067 × 800 pixel datasets to perform the extended experiment. Each image was cropped into multiple 200 × 200 pixel patches so that the proposed ensemble models could do patch classification, just like the GasHisSDB dataset. The 534 × 400 pixel images were cropped into 4 patches, and the 1067 × 800 pixel images were cropped into 20 patches. Each patch was directly assigned the class label of its corresponding image. The same preprocessing steps, models, and experiment settings as described in Section 3 were used, except the output softmax layers of the models were set to three nodes instead of two, to cater for the three classes in the extended experiment dataset.

As shown in Table 8, our proposed ensemble models had good generalization ability and achieved the highest overall accuracies of 96.47% using Ensemble-MV5 and 97.99% using Ensemble-WA5 and Ensemble-UA5 for the 534 × 400 pixel and 1067 × 800 pixel datasets, respectively. The results were substantially better than the performance reported by [55]. For the 534 × 400 pixel dataset, the accuracy improvements are 17.3–37.54% and

23.29% for single classes and overall accuracy, respectively. For the 1067 × 800 pixel dataset, the accuracy improvements are 8.54–19.17% and 14.84% for single classes and overall accuracy, respectively.

Table 8. Performance of the different deep learning models on the 534 × 400 and 1067 × 800 pixel HICL Larynx datasets. The best-achieved results are bold. For the ensemble learning models, WA stands for weighted averaging; UA stands for unweighted averaging, and MV stands for majority voting, and the 5 at the end of the ensemble models refer to the top 5 base models.

		534 × 400 Pixels Dataset				**1067 × 800 Pixels Dataset**			
Paper	**Model**	**Single Class Accuracy (%)**			**Accuracy (%)**	**Single Class Accuracy (%)**			**Accuracy (%)**
		G1	**G2**	**G3**		**G1**	**G2**	**G3**	
[55]	LPCANet	81.18	74.46	60.42	73.18	81.30	89.40	78.50	83.15
Our models	**Ensemble-WA5**	98.48	89.09	95.92	94.71	**98.28**	**97.94**	**97.67**	**97.99**
	Ensemble-UA5	98.48	89.09	95.92	94.71	**98.28**	**97.94**	**97.67**	**97.99**
	Ensemble-MV5	**98.48**	**92.73**	**97.96**	**96.47**	97.99	97.94	97.67	97.88

More complete performance comparison using other evaluation metrics are available in the Supplementary Materials (Tables S4 and S5); our proposed ensemble models easily beat the best reported results in the literature. All these demonstrated the ability and generalization of our proposed ensemble models to handle different histopathology datasets of different organ origins, different staining methods, and multi-class classification tasks.

5.4. Limitations of Our Proposed Study

Although our proposed ensemble models managed to obtain state-of-the-art results in the main and extended dataset, there are several limitations of the work. Firstly, the computational costs are higher due to multiple base models being required to perform the ensemble learning. This is less of an issue compared to the excellent performance offered by the proposed work because histopathological analysis is still manually done by a pathologist in clinical practice at the moment. In addition, graphical processing units (GPUs) can be used to run the base models in parallel to minimize the processing time needed for ensemble learning.

Secondly, the pre-trained weights are transferred from the ImageNet natural image dataset. These are sub-optimal for the histopathology tasks because the image features of the two datasets are quite different, potentially limiting the base model performance. A possible way to address this issue is to fine-tune the base models on another histopathology dataset prior to the target histopathology dataset. By doing so, the networks would have more optimal initial weights on the target histopathology dataset. This should contribute to an improved histopathological detection.

6. Conclusions

In this paper, deep ensemble learning models based on transfer learning of several pre-trained networks such as MobileNet, DenseNet, EfficientNet, InceptionV3, and Xception were developed for gastric cancer detection. It was found that ensemble learning based on the top 5 base models managed to achieve state-of-the-art detection accuracy; this ranged from 97.72 to 99.20% when the image resolution of the histopathological images changed from 80 × 80 pixels to 160 × 160 pixels. The experimental results demonstrated that ensemble models could extract sufficient important features from a smaller patch size yet achieve promising performance. This could lead to lower specifications of the digital scanner, data storage, and computational server required in the histopathology tasks, leading to faster gastric cancer detection and subsequently higher survival rate. In the future, we plan to use GPUs to speed up the processing and to fine-tune the base models on another histopathology dataset first to obtain more optimal initial weights to further improve the performance of gastric cancer detection.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/diagnostics13101793/s1, Table S1: Complete model validation performance tables on the GasHisSDB gastric dataset 80-pixel sub-database, Table S2: Complete model validation performance tables on the GasHisSDB gastric dataset 120-pixel sub-database, Table S3: Complete model validation performance tables on the GasHisSDB gastric dataset 160-pixel sub-database, Figure S1: HICL dataset samples and summary, Tables S4–S6: Complete Performance Evaluation on the 80-, 120-, and 160-pixel sub-databases, Table S7: HICL dataset distribution, Table S8: Complete model test performance tables on the HICL larynx 534 × 400 pixel dataset, Table S9: Complete model test performance tables on the HICL larynx 1067 × 800 pixel dataset.

Author Contributions: Conceptualization, M.P.Y., Y.C.H. and Y.K.T.; methodology, M.P.Y., K.W.L., Y.L.L. and C.-H.G.; formal analysis, M.P.Y., W.-S.Y. and K.W.L.; writing—original draft preparation, M.P.Y., Y.L.L. and C.-H.G.; writing—review and editing, Y.C.H., W.-S.Y. and Y.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to gratefully acknowledge the support of Fundamental Research Grant Scheme (Ref: FRGS/1/2019/ICT04/UTAR/02/1, vote account no: 8073/Y01) and Universiti Tunku Abdul Rahman (UTAR) for providing the UTAR Strategy Research Fund (IPSR/RMC/UTARSF-SR/PROJECT 2021-C1/001) for this work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GasHisSDB dataset is openly available at https://gitee.com/neuhwm/GasHisSDB (accessed on 13 May 2023). The HICL dataset is available upon request at http://medisp.bme.teiath.gr/hicl/ (accessed on 13 May 2023).

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] [PubMed]
2. Shah, M.A.; Khanin, R.; Tang, L.; Janjigian, Y.Y.; Klimstra, D.S.; Gerdes, H.; Kelsen, D.P. Molecular Classification of Gastric Cancer: A New Paradigm. *Clin. Cancer Res.* **2011**, *17*, 2693–2701. [CrossRef] [PubMed]
3. Yasui, W.; Oue, N.; Kuniyasu, H.; Ito, R.; Tahara, E.; Yokozaki, H. Molecular Diagnosis of Gastric Cancer: Present and Future. *Gastric Cancer* **2001**, *4*, 113–121. [CrossRef] [PubMed]
4. Fu, X.; Liu, S.; Li, C.; Sun, J. MCLNet: An Multidimensional Convolutional Lightweight Network for Gastric Histopathology Image Classification. *Biomed. Signal Process. Control* **2023**, *80*, 104319. [CrossRef]
5. Song, Z.; Zou, S.; Zhou, W.; Huang, Y.; Shao, L.; Yuan, J.; Gou, X.; Jin, W.; Wang, Z.; Chen, X.; et al. Clinically Applicable Histopathological Diagnosis System for Gastric Cancer Detection Using Deep Learning. *Nat. Commun.* **2020**, *11*, 4294. [CrossRef]
6. Kato, M.; Nishida, T.; Yamamoto, K.; Hayashi, S.; Kitamura, S.; Yabuta, T.; Yoshio, T.; Nakamura, T.; Komori, M.; Kawai, N.; 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* **2013**, *62*, 1425–1432. [CrossRef]
7. Satolli, M.A. Gastric Cancer: The Times They Are a-Changin'. *World J. Gastrointest. Oncol.* **2015**, *7*, 303. [CrossRef]
8. Bria, E.; De Manzoni, G.; Beghelli, S.; Tomezzoli, A.; Barbi, S.; Di Gregorio, C.; Scardoni, M.; Amato, E.; Frizziero, M.; Sperduti, I.; et al. A Clinical–Biological Risk Stratification Model for Resected Gastric Cancer: Prognostic Impact of Her2, Fhit, and APC Expression Status. *Ann. Oncol.* **2013**, *24*, 693–701. [CrossRef]
9. Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; van Ginneken, B.; Karssemeijer, N.; Litjens, G.; van der Laak, J.A.W.M.; Hermsen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA* **2017**, *318*, 2199. [CrossRef]
10. Hameed, Z.; Zahia, S.; Garcia-Zapirain, B.; Javier Aguirre, J.; María Vanegas, A. Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models. *Sensors* **2020**, *20*, 4373. [CrossRef]
11. Sharma, H.; Zerbe, N.; Klempert, I.; Hellwich, O.; Hufnagl, P. Deep Convolutional Neural Networks for Automatic Classification of Gastric Carcinoma Using Whole Slide Images in Digital Histopathology. *Comput. Med. Imaging Graph.* **2017**, *61*, 2–13. [CrossRef]

12. Korkmaz, S.A.; Akcicek, A.; Binol, H.; Korkmaz, M.F. Recognition of the Stomach Cancer Images with Probabilistic HOG Feature Vector Histograms by Using HOG Features. In Proceedings of the 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 14–16 September 2017; IEEE: Piscataway, NJ, USA; pp. 000339–000342.
13. Korkmaz, S.A.; Binol, H.; Akcicek, A.; Korkmaz, M.F. A Expert System for Stomach Cancer Images with Artificial Neural Network by Using HOG Features and Linear Discriminant Analysis: HOG_LDA_ANN. In Proceedings of the 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 14–16 September 2017; IEEE: Piscataway, NJ, USA; pp. 000327–000332.
14. Sharma, H.; Zerbe, N.; Heim, D.; Wienert, S.; Behrens, H.-M.; Hellwich, O.; Hufnagl, P. A Multi-Resolution Approach for Combining Visual Information Using Nuclei Segmentation and Classification in Histopathological Images. In Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Berlin, Germany, 11–14 March 2015; SCITEPRESS—Science and Technology Publications: Setúbal, Portugal, 2015; pp. 37–46.
15. Sharma, H.; Zerbe, N.; Boger, C.; Wienert, S.; Hellwich, O.; Hufnagl, P. A Comparative Study of Cell Nuclei Attributed Relational Graphs for Knowledge Description and Categorization in Histopathological Gastric Cancer Whole Slide Images. In Proceedings of the 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Thessaloniki, Greece, 22–24 June 2017; IEEE: Piscataway, NJ, USA; pp. 61–66.
16. Liu, B.; Zhang, M.; Guo, T.; Cheng, Y. Classification of Gastric Slices Based on Deep Learning and Sparse Representation. In Proceedings of the 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; IEEE: Piscataway, NJ, USA; pp. 1825–1829.
17. Litjens, G.; Sánchez, C.I.; Timofeeva, N.; Hermsen, M.; Nagtegaal, I.; Kovacs, I.; Hulsbergen-van de Kaa, C.; Bult, P.; van Ginneken, B.; van der Laak, J. Deep Learning as a Tool for Increased Accuracy and Efficiency of Histopathological Diagnosis. *Sci. Rep.* **2016**, *6*, 26286. [CrossRef]
18. Bayramoglu, N.; Kannala, J.; Heikkila, J. Deep Learning for Magnification Independent Breast Cancer Histopathology Image Classification. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: Piscataway, NJ, USA; pp. 2440–2445.
19. Koo, J.C.; Ke, Q.; Hum, Y.C.; Goh, C.H.; Lai, K.W.; Yap, W.-S.; Tee, Y.K. Non-Annotated Renal Histopathological Image Analysis with Deep Ensemble Learning. *Quant. Imaging Med. Surg.* 2023, *in press*.
20. Kong, B.; Wang, X.; Li, Z.; Song, Q.; Zhang, S. Cancer Metastasis Detection via Spatially Structured Deep Network. In Proceedings of the Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, 25–30 June 2017; Springer International Publishing: Cham, Switzerland, 2017; pp. 236–248.
21. BenTaieb, A.; Hamarneh, G. Predicting Cancer with a Recurrent Visual Attention Model for Histopathology Images. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 129–137.
22. Chen, M.; Zhang, B.; Topatana, W.; Cao, J.; Zhu, H.; Juengpanich, S.; Mao, Q.; Yu, H.; Cai, X. Classification and Mutation Prediction Based on Histopathology H&E Images in Liver Cancer Using Deep Learning. *NPJ Precis. Oncol.* **2020**, *4*, 14. [CrossRef]
23. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and Mutation Prediction from Non–Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* **2018**, *24*, 1559–1567. [CrossRef]
24. Nagpal, K.; Foote, D.; Liu, Y.; Chen, P.-H.C.; Wulczyn, E.; Tan, F.; Olson, N.; Smith, J.L.; Mohtashamian, A.; Wren, J.H.; et al. Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer. *NPJ Digit. Med.* **2019**, *2*, 48. [CrossRef]
25. Bychkov, D.; Linder, N.; Turkki, R.; Nordling, S.; Kovanen, P.E.; Verrill, C.; Walliander, M.; Lundin, M.; Haglund, C.; Lundin, J. Deep Learning Based Tissue Analysis Predicts Outcome in Colorectal Cancer. *Sci. Rep.* **2018**, *8*, 3395. [CrossRef]
26. Bulten, W.; Pinckaers, H.; van Boven, H.; Vink, R.; de Bel, T.; van Ginneken, B.; van der Laak, J.; Hulsbergen-van de Kaa, C.; Litjens, G. Automated Deep-Learning System for Gleason Grading of Prostate Cancer Using Biopsies: A Diagnostic Study. *Lancet Oncol.* **2020**, *21*, 233–241. [CrossRef]
27. Koo, J.C.; Hum, Y.C.; Lai, K.W.; Yap, W.-S.; Manickam, S.; Tee, Y.K. Deep Machine Learning Histopathological Image Analysis for Renal Cancer Detection. In Proceedings of the 8th International Conference on Computing and Artificial Intelligence, Tianjin, China, 18–21 March 2022; ACM: New York, NY, USA, 2022; pp. 657–663.
28. Voon, W.; Hum, Y.C.; Tee, Y.K.; Yap, W.-S.; Salim, M.I.M.; Tan, T.S.; Mokayed, H.; Lai, K.W. Performance Analysis of Seven Convolutional Neural Networks (CNNs) with Transfer Learning for Invasive Ductal Carcinoma (IDC) Grading in Breast Histopathological Images. *Sci. Rep.* **2022**, *12*, 19200. [CrossRef]
29. Li, Y.; Li, X.; Xie, X.; Shen, L. Deep Learning Based Gastric Cancer Identification. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; IEEE: Piscataway, NJ, USA; pp. 182–185.
30. Iizuka, O.; Kanavati, F.; Kato, K.; Rambeau, M.; Arihiro, K.; Tsuneki, M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci. Rep.* **2020**, *10*, 1504. [CrossRef]
31. Kloeckner, J.; Sansonowicz, T.K.; Rodrigues, Á.L.; Nunes, T.W.N. Multi-Categorical Classification Using Deep Learning Applied to the Diagnosis of Gastric Cancer. *J. Bras. Patol. Med. Lab.* **2020**, *56*, 1–8. [CrossRef]

32. Hu, W.; Li, C.; Li, X.; Rahaman, M.M.; Ma, J.; Zhang, Y.; Chen, H.; Liu, W.; Sun, C.; Yao, Y.; et al. GasHisSDB: A New Gastric Histopathology Image Dataset for Computer Aided Diagnosis of Gastric Cancer. *Comput. Biol. Med.* **2022**, *142*, 105207. [CrossRef] [PubMed]
33. Doyle, S.; Agner, S.; Madabhushi, A.; Feldman, M.; Tomaszewski, J. Automated Grading of Breast Cancer Histopathology Using Spectral Clusteringwith Textural and Architectural Image Features. In Proceedings of the 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Paris, France, 14–17 May 2008; IEEE: Piscataway, NJ, USA; pp. 496–499.
34. Kather, J.N.; Weis, C.-A.; Bianconi, F.; Melchers, S.M.; Schad, L.R.; Gaiser, T.; Marx, A.; Zöllner, F.G. Multi-Class Texture Analysis in Colorectal Cancer Histology. *Sci. Rep.* **2016**, *6*, 27988. [CrossRef] [PubMed]
35. Sharma, S.; Mehra, R. Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—A Comparative Insight. *J. Digit. Imaging* **2020**, *33*, 632–654. [CrossRef] [PubMed]
36. Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast Cancer Multi-Classification from Histopathological Images with Structured Deep Learning Model. *Sci. Rep.* **2017**, *7*, 4172. [CrossRef]
37. Ameh Joseph, A.; Abdullahi, M.; Junaidu, S.B.; Hassan Ibrahim, H.; Chiroma, H. Improved Multi-Classification of Breast Cancer Histopathological Images Using Handcrafted Features and Deep Neural Network (Dense Layer). *Intell. Syst. Appl.* **2022**, *14*, 200066. [CrossRef]
38. Al-Haija, Q.A.; Adebanjo, A. Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network. In Proceedings of the 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 9–12 September 2020; IEEE: Piscataway, NJ, USA; pp. 1–7.
39. Shallu, S.; Mehra, R. Breast Cancer Histology Images Classification: Training from Scratch or Transfer Learning? *ICT Express* **2018**, *4*, 247–254. [CrossRef]
40. Celik, Y.; Talo, M.; Yildirim, O.; Karabatak, M.; Acharya, U.R. Automated Invasive Ductal Carcinoma Detection Based Using Deep Transfer Learning with Whole-Slide Images. *Pattern Recognit. Lett.* **2020**, *133*, 232–239. [CrossRef]
41. Ghosh, S.; Bandyopadhyay, A.; Sahay, S.; Ghosh, R.; Kundu, I.; Santosh, K.C. Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104202. [CrossRef]
42. Zheng, Y.; Li, C.; Zhou, X.; Chen, H.; Xu, H.; Li, Y.; Zhang, H.; Li, X.; Sun, H.; Huang, X.; et al. Application of Transfer Learning and Ensemble Learning in Image-Level Classification for Breast Histopathology. *Intell. Med.* **2022**. [CrossRef]
43. Paladini, E.; Vantaggiato, E.; Bougourzi, F.; Distante, C.; Hadid, A.; Taleb-Ahmed, A. Two Ensemble-CNN Approaches for Colorectal Cancer Tissue Type Classification. *J. Imaging* **2021**, *7*, 51. [CrossRef]
44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; IEEE: Piscataway, NJ, USA; pp. 2818–2826.
45. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA; pp. 1800–1807.
46. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA; pp. 2261–2269.
47. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PLMR: La Jolla, CA, USA, 2019.
48. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
49. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
50. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
51. Springenberg, M.; Frommholz, A.; Wenzel, M.; Weicken, E.; Ma, J.; Strodthoff, N. From CNNs to Vision Transformers—A Comprehensive Evaluation of Deep Learning Models for Histopathology. *arXiv* **2022**, arXiv:2204.05044.
52. Li, S.; Liu, W. LGFFN-GHI: A Local-Global Feature Fuse Network for Gastric Histopathological Image Classification. *J. Comput. Commun.* **2022**, *10*, 91–106. [CrossRef]
53. Ninos, K.; Kostopoulos, S.; Kalatzis, I.; Sidiropoulos, K.; Ravazoula, P.; Sakellaropoulos, G.; Panayiotakis, G.; Economou, G.; Cavouras, D. Microscopy Image Analysis of P63 Immunohistochemically Stained Laryngeal Cancer Lesions for Predicting Patient 5-Year Survival. *Eur. Arch. Oto-Rhino-Laryngol.* **2016**, *273*, 159–168. [CrossRef]

54. Kostopoulos, S.; Ravazoula, P.; Asvestas, P.; Kalatzis, I.; Xenogiannopoulos, G.; Cavouras, D.; Glotsos, D. Development of a Reference Image Collection Library for Histopathology Image Processing, Analysis and Decision Support Systems Research. *J. Digit. Imaging* **2017**, *30*, 287–295. [CrossRef]
55. Zhou, X.; Tang, C.; Huang, P.; Mercaldo, F.; Santone, A.; Shao, Y. LPCANet: Classification of Laryngeal Cancer Histopathological Images Using a CNN with Position Attention and Channel Attention Mechanisms. *Interdiscip. Sci.* **2021**, *13*, 666–682. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.