Gastric Cancer Image Classification: a Comparative Analysis and Feature Fusion Strategies

Andrea Loddo *, Marco Usai , Cecilia Di Ruberto

Posted Date: 6 July 2024

doi: 10.20944/preprints202407.0478.v1

Keywords: Computational Pathology; Histopathological Imaging; Gastric Cancer; Convolutional Neural Networks; Machine Learning; Deep Learning; Feature Extraction; Feature Combination

 doi:10.20944/preprints202407.0478.v1

Article

Gastric Cancer Image Classification: a Comparative Analysis and Feature Fusion Strategies

Andrea Loddo *, Marco Usai and Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy; m.usai84@studenti.unica.it (M.U.); cecilia.dir@unica.it (C.D.R.)

* Correspondence: andrea.loddo@unica.it

Abstract: Gastric cancer is the fifth most common and fourth deadliest cancer worldwide, with a bleak 5-year survival rate of about 20%. Despite significant research into its pathobiology, prognostic predictability remains insufficient due to pathologists' heavy workloads and the potential for diagnostic errors. Consequently, there is a pressing need for automated and precise histopathological diagnostic tools. This study leverages Machine Learning and Deep Learning techniques to classify histopathological images into healthy and cancerous categories. By utilizing both handcrafted and deep features and shallow learning classifiers on the GasHisSDB dataset, we conduct a comparative analysis to identify the most effective combinations of features and classifiers for differentiating normal from abnormal histopathological images without employing fine-tuning strategies. Our methodology achieves an accuracy of 95% with the SVM classifier, underscoring the effectiveness of feature fusion strategies. Additionally, cross-magnification experiments produced promising results with accuracies close to 80% and 90% when testing the models on unseen testing images with different resolutions.

Keywords: computational pathology; histopathological imaging; gastric cancer; convolutional neural networks; machine learning; deep learning; feature extraction; feature combination

1. Introduction

Gastric cancer is the fifth most prevalent cancer globally and the fourth leading cause of cancer-related deaths, with a global 5-year survival rate hovering around 20%. Despite significant research into the disease's pathobiology, predicting its progression remains difficult, contributing to the persistently low survival rates. Furthermore, medical diagnostics' intricate and time-consuming nature can lead to missing critical details during microscopic examinations, potentially resulting in misdiagnoses [1,2].

Creating computational tools that can automatically and accurately perform histopathological diagnoses is essential to addressing these challenges. Recent advancements in computer technology, especially in Machine Learning (ML) and Deep Learning (DL), have enabled notable progress. This paper investigates classifying pathology images into two categories: healthy cells and tumor cells. It employs various classifiers, extracting specific handcrafted (HC) features, and utilizes different Convolutional Neural Network (CNN) architectures to derive deep features [2–9].

In this study, we leverage the GasHisSDB dataset to evaluate the accuracy of shallow learning classifiers in classifying pathology images using HC and deep features. Notably, we employ the deep features generated by pre-trained CNNs without applying specific optimization or fine-tuning methods. Additionally, we explored several feature fusion techniques to determine how effectively combining features can enhance classification performance. Furthermore, we perform a cross-magnification experiment to assess the impact of different image resolutions within the dataset.

The contributions of this work are the following:

- we proposed a comparative analysis of various HC and deep features across four different ML classifiers to identify the most stable and high-performing feature-classifier pairs for classifying gastric cancer histopathological images and distinguishing between normal and abnormal cells;
- we explored and analyzed various feature fusion techniques to determine their effectiveness in enhancing classification accuracy in the task at hand;

 doi:10.20944/preprints202407.0478.v1

- we conducted a cross-magnification experiment to evaluate the impact of different image resolutions on classification performance, providing insights into the efficacy of utilizing multiple magnifications in pathology image analysis;
- we thoroughly evaluated the GasHisSDB dataset and compared our results with the state-of-the-art.

The rest of the manuscript is organized as follows. Section 2 reviews existing literature to contextualize our research within the field. Section 3 outlines the dataset and methodologies employed in this study. Section 4 presents the findings of our experiments, highlighting the performance of different feature categories, various classifiers, and feature combinations. Section 5 offers an in-depth analysis of our results, comparing them with previous studies and exploring their implications. Finally, the "Conclusion" summarizes our contributions and suggests directions for future research.

2. Related Work

Early detection and accurate diagnosis of gastric cancer (GC) is crucial, as patients with early-stage gastric cancer (EGC) have a much higher 5-year survival rate of 70-90% compared to only 10-30% for advanced gastric cancer (AGC) [3]. However, the accuracy of standard white-light endoscopy for detecting EGC is limited to 70-80%, heavily relying on the expertise of the endoscopist [3]. In recent years, researchers have increasingly explored the use of Computer Vision (CV) and DL techniques to assist in detecting and classifying gastric cancer from endoscopic and pathological images [10].

One of the first studies in this area was by Hirasawa et al., who developed a novel CNN for detecting and recognizing gastric cancer in video images [4]. Similarly, Yoon et al.developed an optimized CNN model for EGC detection and prediction [5].

Beyond endoscopic image analysis, researchers have also explored the use of CV techniques for gastric cancer classification using pathological images. For instance, Zhao et al.conducted a systematic review on the application of CNNs for the identification of gastric cancer [6]. They found that a total of 27 studies had used CNN-based models for gastric cancer detection, classification, segmentation, and margin delineation from various medical imaging modalities, including endoscopy and pathology.

The reported accuracy of the CNN-based systems ranged from 77.3% to 98.7%, demonstrating the strong potential of these techniques for assisting clinicians in the diagnosis of gastric cancer [6].

One notable study in this domain was by Xie et al., who developed an optimized GoogleNet model for the diagnosis of gastric cancer pathological images [7]. Their improved model, which combined the strengths of two network structures, achieved a sensitivity of 97.61% and a specificity of 99.47% in recognizing gastric cancer pathological sections [7].

In this context, Hu et al.proposed a comprehensive dataset, named GasHisSDB, with 245,196 sub-sized gastric histopathology images labelled as normal or gastric cancer, which were derived from 600 whole slide images (WSIs) [2]. It was introduced to overcome the limitations of the existing datasets, particularly their small sample sizes [2,11].

Several follow-up studies have used the GasHisSDB dataset, starting from its proposal, where Hu et al.evaluated the performance of various ML and DL models [8], while several authors proposed optimized approaches to accomplish this task. For instance, Yong et al.proposed an ensemble DL approach based on EfficientNetB0, EfficientNetB1, DenseNet-121, DenseNet169, and MobileNet [9] whereas Li et al.introduced a lightweight gated fully-fused network (LGFFN) with a gated hybrid input (GHI) module. The LGFFN-GHI comprises two main components: feature extraction and classification modules. The feature extraction module uses a cross-attention mechanism to fuse features from different scales. The classification module then takes the fused features and outputs the final classification prediction [12]. In addition, Fu et al.proposed MCLNet, a multidimensional CNN based on ShuffleNet. It extracts the correlation features between pixels in an image by one-dimensional convolution to achieve pixel-level and patch-level feature interaction.

Overall, the reviewed studies demonstrate significant progress in applying CV and DL techniques for gastric cancer classification from endoscopic and pathological images. In this context, our study aims to advance the classification of gastric cancer using histopathologic images. Our primary objective

 doi:10.20944/preprints202407.0478.v1

is to propose a robust system that does not rely on ad-hoc adjustments or fine-tuning. By leveraging features from non-optimized yet generic methods, we explore the feasibility of offering a generalizable solution that performs consistently across various magnifications and datasets.

Considering that the performance of these systems can be influenced by factors such as the size and diversity of the training datasets, the specific CNN architectures employed, and the clinical context in which they are deployed, we propose a pipeline specifically based on HC features and features extracted by pre-trained CNNs.

3. Materials and Methods

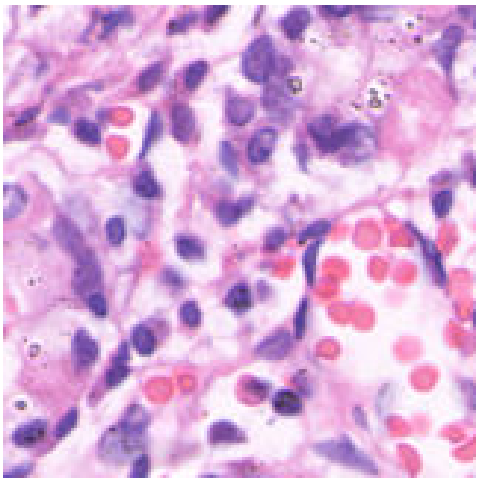
This section provides details of the components used in our study. We begin with an overview of the dataset in Section 3.1, detailing its composition and relevance to our research objectives. Following this, in Section 3.2, we present both feature extraction methods employed, HC and deep, whereas in Section 3.3, we describe the classification methods applied. Finally, we discuss the performance evaluation measures in Section 3.4.

3.1. Dataset

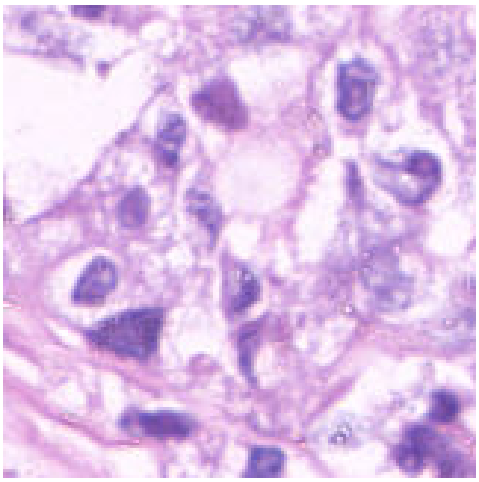
The GasHisSDB dataset is a publicly available gastric histopathology image dataset [2]. It contains a total of 245,196 sub-sized gastric histopathology images, which were derived from 600 WSIs, stained with H&E, of 2048 $\times$ 2048 pixels. The images were scanned using a NewUsbCamera and digitized at 20$\times$ magnification. Two experienced pathologists from Liaoning Cancer Hospital and Institute provided the labels, classifying the images as either normal or abnormal (gastric cancer). A normal image is characterized by the absence of cancerous regions, reflecting typical microscopic cell observations. In contrast, an image is labeled as abnormal when approximately 50% of its area is occupied by cancerous regions [2]. The dataset is divided into three image sub-databases, each of them containing images with specific resolutions: 160 $\times$ 160 (S-A), 120 $\times$ 120 (S-B), and 80 $\times$ 80 (S-C) pixels. The distribution of the dataset images is summarized in Table 1, while Figure 1 shows two image samples.

Table 1. Description of GasHisSDB with details on its subdivision and number of images per class.

Sub-database	Size	Abnormal	Normal
S-A	160 $\times$ 160 pixels	13,124	20,160
S-B	120 $\times$ 120 pixels	24,801	40,460
S-C	80 $\times$ 80 pixels	59,151	87,500
Total		97,076	148,120



(a) Normal class image.



(b) Abnormal class image.

Figure 1. Sample images from the GasHisSDB dataset.

 doi:10.20944/preprints202407.0478.v1

3.2. Feature Extraction Methods

Features derived from images include a wide array of descriptors designed to capture morphological, pixel-level, and textural information, denoted as handcrafted features. As noted by [13], HC features can be broadly categorized into three main groups: invariant moments, texture features, and color features. To them, we have added a set of deep features, i.e., features obtained by the activations of off-the-shelf CNNs. In the following, we present a brief summary of each category along with the specific descriptors utilized.

3.2.1. Invariant Moments

An image moment is a weighted average of pixel intensities in an image used to extract specific properties. Moments are crucial in image analysis and pattern recognition, helping to characterize segmented objects [14]. This study employs three distinct types of moments: Zernike, Legendre, and Chebyshev. A concise overview of these moment types follows.

Chebyshev Moments (CH) constitute a class of discrete orthogonal moments [15], based on Chebyshev polynomials [16] with the maximum possible leading coefficient constrained by an absolute value of 1 within the interval $[-1, 1]$. This study used the first- and second-order moments, denoted as CH_1 and CH_2, respectively. Both moments were calculated to the fifth order.

Second-order Legendre Moments (LM) are a type of continuous orthogonal moments that can be used for image analysis. LM capture information about the shape and orientation of an image. They are calculated using the second-order Legendre polynomials, which are orthogonal over the interval $[-1, 1]$ [17,18]; they can capture and represent objects' shape and spatial characteristics within an image. In our analysis, we extracted the LM of order 5.

Zernike Moments (ZM) are a type of continuous orthogonal moments that are defined over the unit circle [18]. They are calculated using Zernike polynomials, which form an orthogonal basis [19]. In this study, we extracted the ZM of order 5 with a repetition of 5.

3.2.2. Texture Features

Texture serves as a visual feature indicative of homogeneity within an image. It reveals the organization and arrangement of surface structures exhibiting gradual or periodic variations. Rather than relying on individual pixel characteristics, texture analysis requires statistical calculations over regions encompassing multiple pixels [20]. The texture is characterized by the gray-level distribution of pixels and their surrounding spatial neighbors, encapsulating local texture information. Additionally, global texture information is determined by the extent of repetition of this local texture information. For the sake of this work, we have considered two widely employed methods, now described.

Rotation-Invariant Haralick (HAR) Features. Thirteen HAR features were extracted from the Gray Level Co-occurrence Matrix (GLCM) and then transformed into rotation-invariant features [21]. To achieve rotation invariance, four variations of the GLCM were calculated, each with a distance parameter $d = 1$ and angular orientations $\theta = [0°, 45°, 90°, 135°]$.

Local Binary Pattern (LBP) is a powerful method for capturing the texture and patterns in an image, as described by [22]. In this study, we computed the histogram of the LBP and transformed it into a rotation-invariant form [23]. This histogram was then extracted and used as the feature vector. The LBP map was generated within a neighborhood defined by a radius of $r = 1$ and eight neighbors $(n = 8)$.

3.2.3. Color Features

Histograms are the most widely employed method for characterizing the color properties of images since they effectively represent the global color distribution within an image, indicating their proportion. The descriptors that can be extracted from the histogram are invariant to image rotation, translation, and scaling changes. However, they have a significant limitation in that they cannot describe the local distribution of colors, the spatial location of each color, or specific objects within the

 doi:10.20944/preprints202407.0478.v1

image [24]. In this study, these descriptors were calculated from images that underwent a conversion to grayscale, streamlining the process of analysis and computation.

Histogram (Hist) Features. From the histogram, which characterizes the overall color distribution within the image, we derived seven statistical descriptors: mean, standard deviation, smoothness, skewness, kurtosis, uniformity, and entropy.

Autocorrelogram (AC). The AC captures the spatial correlation of colors within an image. It is a restricted version of the more general color correlogram, considering only the spatial correlation between pixels of the same color [25]. Specifically, the color autocorrelogram calculates the probability that a pixel of a given color will be found at a certain distance, d, away from another pixel of the same color. In our research, we considered four discrete distances: $d = 1, 2, 3, 4$. The four resulting probability vectors are concatenated to form a comprehensive feature vector.

Haar-like (Haar) Features. The key idea behind these features is to calculate the difference in the sum of pixel intensities across rectangular regions in an image. This allows detecting edges, lines, and center-surround features that indicate the presence of an object [26].

Table 2. Employed CNNs details including reference paper, number of trainable parameters in millions, input shape, feature extraction layer, and related feature vector size.

Reference	Parameters (M)	Input shape	Feature layer	# Features
AlexNet [27]	60	224 × 224	Pen. FC	4,096
DarkNet-19 [28]	20.8	224 × 224	Conv19	1,000
DarkNet-53 [29]	20.8	224 × 224	Conv53	1,000
DenseNet-201 [30]	25.6	224 × 224	Avg. Pool	1,920
EfficientNetB0 [31]	5.3	224 × 224	Avg. Pool	1,280
Inception-v3 [32]	21.8	299 × 299	Last FC	1,000
Inception-ResNet-v2 [33]	55	299 × 299	Avg. pool	1,536
ResNet-18 [34]	11.7	224 × 224	Pool5	512
ResNet-50 [34]	26	224 × 224	Avg. Pool	1,024
ResNet-101 [34]	44.6	224 × 224	Pool5	1,024
VGG19 [35]	144	224 × 224	Pen. FC	4,096
XceptionNet [36]	22.9	299 × 299	Avg. Pool	2,048

3.2.4. Deep Features

With deep features, we refer to the characteristics of an image derived from the CNN activations since they have proven to be a potent strategy for enhancing the predictive power of classifiers [37]. These deep features were extracted from off-the-shelf CNN architectures pretrained on the well-known natural image dataset ImageNet [38].

Specifically, depending on the architecture, deep features were extracted from one of the following layers: i) the penultimate layer, ii) the final fully connected layer, or iii) the last pooling layer. This approach ensures extracting features encapsulating the network's learned global knowledge. Notably, the fine-tuning strategy for the classification phase was not employed to maintain the generalization ability of the networks [39,40]. Detailed specifications regarding the selected layers for feature extraction, input dimensions, and the count of trainable parameters for each CNN model are outlined in Table 2, while a brief explanation of the CNNs employed is now provided.

AlexNet consists of a sequence of convolutional and max pooling layers, culminating in three fully connected layers [27]. With only five convolutional layers, it represents the most shallow architecture used in this study.

DarkNet builds upon the established principles of inception and batch normalization. This study employs two specific versions of DarkNet, incorporating 19 [28] and 53 [29] convolutional layers. These configurations form the foundational network for the You Only Look Once object detection method.

 doi:10.20944/preprints202407.0478.v1

DenseNet, proposed by Huang et al. [30], addresses the typical CNN characteristic of having layers equal to the number of connections. Specifically, the number of connections is $L(L+1)/2$, where L denotes the number of layers. Each layer's input comprises the output from all preceding layers, which then serves as the input for the subsequent layer.
EfficientNet stands out for its uniform and efficient scaling of network width, depth, and resolution through compound scaling. Proposed by Tan et al. [31], this study employs EfficientNetB0 version.
Inception-v3 uses the inception layer concept by incorporating factorized, smaller, and asymmetric convolutions [32]. Inception models are notable for their multi-branch architectures, combining filters of various sizes integrated through concatenation within each branch.
Inception-ResNet-v2 merges the strengths of ResNet and Inception architectures [33]. The Inception-ResNet block combines variously sized convolutional filters with residual connections, featuring four max-pooling layers and 160 convolutional layers.
ResNet refers to a family of deep architectures that use residual learning [34]. These architectures integrate skip-connections or recurrent units to link blocks of convolutional and pooling layers, with each block followed by batch normalization [41]. This study employs three ResNet variants: ResNet-18, ResNet-50, and ResNet-101, with the numbers indicating the respective network depths.
VGG comprises a series of convolutional layers followed by max-pooling, which enhances its deep representation capabilities [35]. This study uses VGG19, featuring 19 layers.
XceptionNet extends the *Inception* architecture by employing depth-wise separable convolutions to improve efficiency and reduce parameter count. This approach aims to capture complex feature dependencies by focusing on cross-channel correlations [36].

3.3. Classification Methods

After feature extraction, HC and deep features served as inputs for four classical ML algorithms to classify GasHisSDB. Here is a brief overview of these classifiers.
Decision Tree (DT) is a hierarchical data structure used for prediction. Each internal node represents a feature, with branches denoting possible feature values and leaves representing different categories. The algorithm optimizes this structure by pruning nodes that minimally contribute to category separation, thereby merging instances at higher levels. Classification is achieved by tracing the path from the root to a leaf node [42].
k-Nearest Neighbor (kNN): The kNN classifier categorizes observations by considering the classes of the k training examples nearest to the observation in question. This method employs a local strategy for classification, leveraging the proximity of neighboring instances to determine the class [43].
Support Vector Machine (SVM): SVM differentiates categories by mapping examples to opposite sides of a decision boundary. The one-vs-rest approach is employed for multiclass problems, training individual classifiers to distinguish each class from all others [44].
Random Forest (RF): This algorithm aggregates predictions from multiple decision trees, each constructed from random subsets of features and examples. By fostering diversity among the trees, this ensemble method enhances model robustness, improving resilience against data imbalance and mitigating overfitting. The use of 100 trees specifically enhances the random forest's predictive accuracy [45].

3.4. Performance Evaluation Measures

In evaluating the performance of a binary classifier on a dataset, each instance is classified as either negative or positive based on the classifier's predictions. The result of this classification, when compared to the actual target value, determines the following performance measures:

- **True Negatives (TN)**: instances correctly predicted as negative.
- **False Positives (FP)**: instances incorrectly predicted as positive.
- **False Negatives (FN)**: instances incorrectly predicted as negative.
- **True Positives (TP)**: instances correctly predicted as positive.

 doi:10.20944/preprints202407.0478.v1

We assess the classifier's performance using several measures specifically defined for binary classification tasks, as detailed below.

- **Accuracy (A)**. It is the ratio of correct predictions to the total number of predictions:

$$\mathrm{A} = \frac{TP + TN}{TP + FN + FP + TN}$$

- **Precision (P)** is the ratio of TPs to the sum of TPs and FPs, indicating the classifier's efficiency in predicting positive instances:

$$\mathrm{P} = \frac{TP}{TP + FP}$$

- **Recall (R)**, also known as sensitivity, is the ratio of TPs to the sum of TPs and FNs:

$$\mathrm{R} = \frac{TP}{TP + FN}$$

- **Specificity (S)** is the ratio of TNs to the sum of TNs and FPs:

$$\mathrm{S} = \frac{TN}{TN + FP}$$

- **F1-score (F1)** is the harmonic mean of P and R, considering both FPs and FNs:

$$F1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

- **Matthews Correlation Coefficient (MCC)** is a comprehensive measure that considers all elements of the confusion matrix (TP, TN, FP, FN). Ranging from -1 to $+1$, it provides a high score only when the classifier performs well in both the positive and negative classes:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **Balanced Accuracy (BACC)** is defined as the mean of specificity and sensitivity:

$$BACC = \frac{\mathrm{S} + \mathrm{R}}{2}$$

4. Experimental Results

In this section, we detail the comprehensive experimental analysis conducted to evaluate the performance of various feature extraction and classification techniques on the GasHisSDB dataset. Section 4.1 outlines the experimental setup and implementation details. Following this, Section 4.2 presents the results obtained with HC features, whereas Section 4.3 explores the use of CNN as feature extractors. This is succeeded by Section 4.4, where we discuss the outcomes of combining HC and deep features to enhance classification accuracy. To ensure robustness across different magnifications, Section 4.5 evaluates the consistency and reliability of our methods when applied to images at various magnification levels. Finally, Section 4.6 provides a critical analysis of our results in the context of existing research.

4.1. Experimental Setup

The experiments were performed on a workstation with an Intel(R) Core(TM) i9-8950HK @ 2.90GHz CPU, 32 GB of RAM, and an NVIDIA GTX1050 Ti GPU with 4GB of memory. MATLAB R2021b was used for all implementations and experimental evaluations.

This study deliberately did not use image augmentations to concentrate on extracting pure features from the original images. Moreover, we used Euclidean distance as a distance measure for

 doi:10.20944/preprints202407.0478.v1

kNN with $k = 1$; note that with $k = 1$, no voting strategy is required. In addition, the SVM kernel function uses a linear kernel, and the number of DTs composing the RF has been set to 100.

Finally, a 5-fold cross-validation (CVal) approach was employed for the testing strategy. This method ensures statistical reliability by repeatedly training and testing the same dataset. Specifically, the dataset is divided into 80% for training and 20% for testing at each iteration.

4.2. HC Features Performance

The outcomes of the classifiers trained with HC features are presented in Tables 3–6. These tables detail the performance results for the DT, kNN, RF, and SVM, respectively.

Table 3. Performance obtained with DT trained with HC features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AC	62.78	69.39	68.97	53.26	69.18	22.20	61.12
Haar	59.70	62.56	83.31	23.43	71.46	8.33	53.37
Hist	68.78	74.49	73.71	61.22	74.10	34.84	67.46
HAR	68.78	74.85	72.99	62.32	73.91	35.10	67.66
LBP	71.22	76.23	76.26	63.47	76.25	39.74	69.87
CH_1	71.11	76.15	76.17	63.35	76.16	39.52	69.76
CH_2	71.05	76.12	76.07	63.35	76.09	39.41	69.71
LM	71.32	76.16	76.64	63.16	76.40	39.87	69.90
ZM	58.06	65.74	64.24	48.57	64.98	12.73	56.40

Table 4. Performance obtained with kNN trained with HC features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AC	57.34	70.04	51.66	66.06	59.46	17.42	58.86
Haar	42.17	56.99	18.40	78.67	27.82	-3.61	48.53
Hist	64.64	71.97	68.15	59.24	70.01	27.07	63.70
HAR	61.06	68.53	66.05	53.41	67.26	19.29	59.73
LBP	69.51	75.32	73.86	62.82	74.58	36.50	68.34
CH_1	66.16	72.41	71.28	58.29	71.84	29.45	64.78
CH_2	65.46	72.09	70.14	58.29	71.10	28.24	64.21
LM	66.32	72.57	71.38	58.55	71.97	29.81	64.97
ZM	57.40	65.03	64.19	46.97	64.60	11.12	55.58

Table 5. Performance obtained with SVM trained with HC features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AC	67.30	69.18	82.99	43.20	75.45	28.71	63.09
Haar	62.18	62.38	94.59	12.38	75.18	12.45	53.49
Hist	44.67	96.29	9.00	99.47	16.47	17.91	54.23
HAR	62.07	73.64	58.21	68.00	65.02	25.64	63.10
LBP	62.64	70.00	67.06	55.85	68.50	22.69	61.46
CH_1	75.92	77.37	85.14	61.75	81.07	48.61	73.45
CH_2	72.50	71.57	90.55	44.76	79.95	40.78	67.66
LM	73.32	72.65	89.73	48.11	80.29	42.61	68.92
ZM	63.84	69.78	71.08	52.72	70.43	23.93	61.90

 doi:10.20944/preprints202407.0478.v1

Table 6. Performance obtained with RF trained with HC features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AC	71.76	72.89	84.97	51.47	78.47	39.09	68.22
Haar	62.48	62.65	94.22	13.71	75.26	13.61	53.97
Hist	73.26	77.24	79.19	64.15	78.20	43.66	71.67
HAR	76.78	78.69	84.55	64.84	81.52	50.63	74.69
LBP	79.57	80.74	87.03	68.11	83.77	56.61	77.57
CH_1	78.11	79.99	85.17	67.28	82.50	53.56	76.22
CH_2	78.07	79.84	85.34	66.90	82.50	53.43	76.12
LM	78.25	80.20	85.09	67.73	82.58	53.87	76.41
ZM	65.19	68.16	79.84	42.70	73.54	24.26	61.27

The performance obtained with the DT (Table 3) shows that the LBP again leads with an accuracy of 71.22% and a BACC of 69.87%. This indicates that despite the DT classifier's simplicity, LBP features can still capture discriminative information effectively. Haar features, however, perform poorly with a balanced accuracy of 53.37%.

With kNN (Table 4), LBP again emerges as the top performer, with an accuracy of 69.51% and an F1 of 74.58%. Conversely, the Haar feature shows poor performance with an accuracy of 42.17% and a balanced accuracy of 48.53%, reaffirming its limitations for this task.

As for the SVM (Table 5), the CH_1 and CH_2 features provide the best accuracy (75.92% and 72.50%, respectively) and balanced accuracy (73.45% and 67.66%, respectively), showing their potential for discriminating between normal and abnormal tissues. Interestingly, while the Hist features achieve high precision (96.29%), it suffers from very low recall (9.00%), leading to a much lower balanced accuracy (54.23%).

Finally, the RF (Table 6) trained with the LBP features achieve the highest accuracy (79.57%), precision (80.74%), and F1 (83.77%). The LBP feature's strong performance across most metrics suggests its effectiveness in capturing essential patterns in histopathological images. Contrastively, the Haar feature again demonstrates a significantly lower accuracy (62.48%) and balanced accuracy (53.97%), indicating its relative ineffectiveness in this context.

The consistency across classifiers underlines LBP features' robustness, even without top-notch performance.

4.3. Deep Features Performance

The results of the classifiers trained using deep features are summarized in Tables 7–10. These tables provide a comprehensive overview of the performance metrics for the DT, kNN, SVM, and RF classifiers.

Table 7. Performance obtained with DT trained with deep features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AlexNet	75.00	79.47	79.19	68.57	79.33	47.72	73.88
DarkNet-19	78.25	82.04	82.04	72.42	82.04	54.46	77.23
DarkNet-53	81.64	84.78	84.95	76.57	84.86	61.55	80.76
DenseNet-201	84.92	87.51	87.60	80.80	87.56	68.42	84.20
EfficientNet B0	78.79	82.62	82.29	73.41	82.46	55.64	77.85
Inception-v3	74.54	79.21	78.60	68.30	78.90	46.81	73.45
Inception-ResNet-v2	73.52	78.02	78.35	66.10	78.18	44.49	72.22
ResNet-18	76.73	81.20	80.13	71.50	80.66	51.46	75.82
ResNet-50	81.30	84.62	84.47	76.42	84.55	60.87	80.45
ResNet-101	80.13	83.67	83.48	74.97	83.58	58.42	79.23
VGG19	76.40	80.97	79.79	71.20	80.37	50.80	75.49
Xception	79.38	83.30	82.49	74.59	82.89	56.94	78.54

 doi:10.20944/preprints202407.0478.v1

Table 8. Performance obtained with kNN trained with deep features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AlexNet	80.85	83.80	84.77	74.82	84.28	59.78	79.80
DarkNet-19	83.99	86.47	87.20	79.05	86.84	66.40	83.13
DarkNet-53	88.25	88.89	92.11	82.32	90.48	75.25	87.22
DenseNet-201	88.21	89.04	91.84	82.63	90.42	75.16	87.23
EfficientNet B0	87.53	89.01	90.60	82.82	89.80	73.79	86.71
Inception-ResNet-v2	76.13	79.89	80.98	68.69	80.43	49.85	74.83
Inception-v3	80.88	83.65	85.04	74.48	84.34	59.80	79.76
ResNet-101	87.89	88.57	91.87	81.79	90.19	74.48	86.83
ResNet-18	84.21	85.89	88.47	77.68	87.16	66.73	83.07
ResNet-50	86.97	88.20	91.09	80.41	89.62	73.19	85.75
VGG19	83.35	85.94	86.94	77.95	86.44	65.88	82.45
Xception	85.15	86.74	89.44	78.95	88.07	70.17	84.20

Table 9. Performance obtained with SVM trained with deep features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AlexNet	72.97	77.49	76.56	65.81	77.02	42.80	71.18
DarkNet-19	77.86	82.20	80.82	71.95	81.50	53.28	76.39
DarkNet-53	82.83	85.84	84.99	76.88	85.41	63.65	80.94
DenseNet-201	86.02	89.01	86.61	81.84	87.79	69.91	84.23
EfficientNet B0	82.84	85.76	85.35	76.96	85.55	63.48	81.15
Inception-v3	73.60	78.88	74.73	67.51	76.75	45.86	71.12
Inception-ResNet-v2	69.87	75.03	71.65	62.78	73.29	35.68	67.21
ResNet-18	77.78	82.17	80.82	71.86	81.49	53.13	76.34
ResNet-50	82.77	86.11	84.34	77.87	85.22	63.43	81.11
ResNet-101	82.52	85.76	84.71	76.81	85.23	63.00	80.76
VGG19	79.60	83.94	81.78	73.70	82.84	57.67	77.74
XceptionNet	82.24	85.75	84.22	76.22	84.97	62.53	80.22

Table 10. Performance obtained with RF trained with deep features. Values are shown in terms of %.

Desc.	A	P	R	S	F1	MCC	BACC
AlexNet	84.02	85.55	88.57	77.03	87.03	66.29	82.80
DarkNet-19	88.30	88.68	92.49	81.87	90.54	75.33	87.18
DarkNet-53	90.30	90.72	93.55	85.30	92.11	79.58	89.42
DenseNet-201	91.93	92.61	94.20	88.46	93.40	83.05	91.33
EfficientNet B0	89.89	89.96	93.77	83.92	91.83	78.71	88.85
Inception-v3	85.52	85.64	91.42	76.46	88.44	69.39	83.94
Inception-ResNet-v2	83.25	84.10	89.21	74.10	86.58	64.55	81.65
ResNet-18	86.99	87.32	91.87	79.50	89.53	72.54	85.68
ResNet-50	89.92	90.12	93.63	84.23	91.84	78.77	88.93
ResNet-101	89.59	89.76	93.48	83.62	91.58	78.07	88.55
VGG19	85.98	86.61	90.92	78.40	88.71	70.41	84.66
Xception	88.58	89.08	92.49	82.59	90.75	75.94	87.54

As shown in Table 10, DT trained with DenseNet-201 features excel, achieving the highest accuracy (84.92%) and balanced accuracy (84.20%). This suggests that even simpler classifiers like DT can benefit significantly from the rich feature representations provided by pre-trained CNNs.

kNN (Table 8) shows superior performance with DenseNet-201 and DarkNet-53, which gained accuracies of 88.21% and 88.25%, respectively, and balanced accuracies of 87.23% and 87.22%. These results further confirm the robustness and high classification potential of deep features extracted from these models.

Even with the SVM (Table 9), DenseNet-201 features achieve the highest accuracy (86.02%) and balanced accuracy (84.23%), followed closely by DarkNet-53 and EfficientNetB0.

 doi:10.20944/preprints202407.0478.v1

Finally, with RF (Table 10), DenseNet-201 achieves the highest accuracy (91.93%) and balanced accuracy (91.33%), indicating its superior feature extraction capability. Other deep features, such as those from DarkNet-53 and ResNet-101, also perform exceptionally well with RF.

This consistent top performance across different classifiers highlights DenseNet-201's strong feature extraction capabilities for histopathological images, even without any fine-tuning strategy.

4.4. Feature Fusion Performance

Despite the significant results achieved from the previous classification, further efforts were made to enhance performance. An additional experiment was conducted to explore the potential of integrating the representative power of both HC and deep features into a feature fusion strategy. More specifically, this experiment focused on three features: LBP, DenseNet-201, and EfficientNetB0, which resulted the best HC and the best two deep features, respectively. They were evaluated with all their possible combinations by using the best three classifiers from the previous stage: DT, SVM, and RF. This integration aimed to leverage their combined strengths for improved performance. The results are shown in Table 11.

Table 11. Performance measures of different classifiers trained with a feature fusion strategy. The classifiers used are DT, SVM, and RF. The strategies compared include combinations of HC and deep features: LBP + DenseNet-201, LBP + EfficientNetB0, DenseNet-201 + EfficientNetB0, and LBP + DenseNet-201 + EfficientNetB0. Values are shown in terms of %.

Strategy	Class.	A	P	R	S	F1	MCC	BACC
LBP + DenseNet-201	DT	88.21	89.04	91.84	82.63	90.42	75.16	87.23
	SVM	94.41	95.14	95.66	92.50	95.40	88.29	94.08
	RF	92.16	92.83	94.35	88.80	93.58	83.53	91.57
LBP + EfficientNetB0	DT	87.55	89.01	90.63	82.82	89.81	73.82	86.72
	SVM	94.05	94.78	95.44	91.92	95.11	87.53	93.68
	RF	89.65	90.19	93.03	84.46	91.59	78.21	88.74
DenseNet-201 + EfficientNetB0	DT	90.30	91.05	93.13	85.94	92.08	79.59	89.54
	SVM	94.89	95.76	95.81	93.49	95.78	89.31	94.65
	RF	91.83	92.31	94.37	87.92	93.33	82.82	91.15
LBP + DenseNet-201 + EfficientNetB0	DT	90.31	91.07	93.13	85.98	92.09	79.63	89.56
	SVM	95.03	95.86	95.93	93.64	95.90	89.59	94.79
	RF	92.26	92.67	94.72	88.50	93.68	83.74	91.61

For the fusion of LBP and DenseNet-201 features, SVM emerged as the most effective classifier with an accuracy of 94.41% and F1 of 95.40%. This performance indicates that SVM, when trained with this fusion of features, is highly reliable. RF also performed robustly with an accuracy of 92.16%, showing strong capability but slightly lagging behind SVM.

In the case of combining LBP with EfficientNetB0 features, SVM again demonstrated superior performance with an accuracy of 94.05%, F1 of 95.11%, MCC of 87.53, and balanced accuracy of 93.68%. This reiterates SVM's effectiveness across different feature combinations. RF showed a notable drop in performance compared to the previous fusion strategy, suggesting that this combination might not be as effective for RF.

The combination of DenseNet-201 and EfficientNetB0 features led to SVM achieving the highest measures overall, with an accuracy of 94.89%, F1-score of 95.78%. This indicates that the deep features from these two CNNs complement each other well, providing rich information for the classifier. RF and DT also performed better with this fusion strategy compared to using HC features, highlighting the benefit of using purely deep features.

Finally, the most complex feature fusion strategy, combining LBP with both DenseNet-201 and EfficientNetB0, resulted in the highest overall performance for SVM, with an accuracy of 95.03%, and F1 of 95.90%. This indicates that the incorporation of both HC and multiple deep features

 doi:10.20944/preprints202407.0478.v1

provides a comprehensive feature set that enhances classification performance. RF also showed its best performance with this fusion strategy, suggesting that the addition of more feature types helps improve its robustness and generalization.

4.5. Cross-Magnification Performance

Tables 12 and 13 detail the performance measures of two cross-magnification experiments conducted using different classifiers and feature fusion strategies on the GasHisSDB dataset. The experiments involved training on a 160 × 160 image sub-database and testing on smaller dimensions (120 × 120 and 80 × 80, respectively).

Table 12. Performance measures of the first cross-magnification experiment. Different classifiers were trained on the 160 × 160 split of the GasHisSDB with a feature fusion strategy and tested on the 120 × 120. The classifiers used were DT, SVM, and RF. Values are shown in terms of %.

Strategy	Class.	A	P	R	S	F1	MCC	BACC
LBP + DenseNet-201	DT	86.69	90.56	87.68	85.08	89.09	72.10	86.38
	SVM	86.41	97.42	80.20	96.53	87.98	74.51	88.37
	RF	89.04	95.12	86.78	92.74	90.76	77.87	89.76
LBP + EfficientNetB0	DT	85.17	88.66	87.25	81.79	87.95	68.71	84.52
	SVM	85.02	96.81	78.42	95.79	86.65	72.04	87.10
	RF	87.38	90.20	89.36	84.15	89.78	73.30	86.76
DenseNet-201 + EfficientNetB0	DT	88.43	91.65	89.50	86.69	90.56	75.66	88.09
	SVM	85.88	98.40	78.50	97.92	87.33	74.19	88.21
	RF	89.55	94.36	88.43	91.37	91.30	78.51	89.90
LBP + DenseNet-201 + EfficientNetB0	DT	88.42	91.64	89.50	86.67	90.55	75.65	88.08
	SVM	85.82	98.45	78.36	97.98	87.26	74.12	88.17
	RF	89.56	94.79	87.99	92.12	91.26	78.67	90.05

Table 13. Performance measures of the second cross-magnification experiment. Different classifiers were trained on the 160 × 160 split of the GasHisSDB with a feature fusion strategy and tested on the 80 × 80. The classifiers used were DT, SVM, and RF. Values are shown in terms of %.

Strategy	Class.	A	P	R	S	F1	MCC	BACC
LBP + DenseNet-201	DT	77.05	86.23	73.23	82.70	79.20	54.89	77.97
	SVM	68.92	96.18	49.89	97.07	65.70	49.83	73.48
	RF	78.89	93.29	69.62	92.60	79.73	61.41	81.11
LBP + EfficientNetB0	DT	63.58	89.19	44.34	92.05	59.23	39.09	68.20
	SVM	54.36	96.04	24.53	98.50	39.07	31.44	61.51
	RF	71.38	92.27	56.79	92.96	70.31	50.63	74.88
DenseNet-201 + EfficientNetB0	DT	74.70	88.69	66.02	87.55	75.69	52.89	76.78
	SVM	59.96	96.81	34.02	98.34	50.34	39.00	66.18
	RF	79.73	92.84	71.54	91.84	80.81	62.39	81.69
LBP + DenseNet-201 + EfficientNetB0	DT	74.70	88.69	66.02	87.55	75.69	52.89	76.78
	SVM	60.31	96.72	34.66	98.26	51.03	39.39	66.46
	RF	78.44	93.88	68.33	93.41	79.09	61.10	80.87

Results of testing on S-B. In the first experiment, the classifiers were evaluated on the 120 × 120 test set. The combination of LBP and DenseNet-201 as features yielded varied results across different classifiers. The RF classifier outperformed others, achieving an accuracy of 89.04%, an F1 of 90.76%, and a balanced accuracy of 89.76%. The SVM also demonstrated strong performance, particularly with a precision of 97.42%, though it lagged in recall compared to RF.

 doi:10.20944/preprints202407.0478.v1

When integrating LBP with EfficientNetB0, the performance metrics showed a slight decline, especially noticeable in the DT classifier, which recorded an accuracy of 85.17%. The RF continued to maintain relatively high performance, albeit slightly lower than with DenseNet-201.

The fusion of DenseNet-201 and EfficientNetB0 features displayed a notable improvement in classifier performance. RF again led the results with an accuracy of 89.55%, an F1 score of 91.30%, and a balanced accuracy of 89.90%. The DT classifier also performed well under this strategy, achieving high precision and recall rates.

Combining all three feature sets (LBP, DenseNet-201, and EfficientNetB0) resulted in marginal improvements across the board. RF achieved the highest accuracy at 89.56%, while the SVM exhibited the highest precision at 98.45%. This comprehensive feature fusion strategy enhanced the robustness and consistency of the classifiers' performance, particularly evident in the balanced accuracy and MCC scores.

Results of testing on S-C. The second experiment, with testing on the 80×80 sub-database, illustrated a greater challenge for the classifiers, reflected in the generally lower performance values. The LBP + DenseNet-201 combination showed that RF remained the most reliable classifier with an accuracy of 78.89% and an F1 score of 79.73%. The SVM, while demonstrating high precision at 96.18%, struggled with recall and balanced accuracy, indicating a possible reliance on the 160×160 pixels image data.

In the LBP + EfficientNetB0 strategy, all classifiers showed decreased performance, with the SVM particularly underperforming in terms of recall and F1. RF again stood out, albeit with lower scores compared to the previous experiment.

The fusion of DenseNet-201 and EfficientNetB0 improved the values slightly, with RF achieving an accuracy of 79.73% and a balanced accuracy of 81.69%. This strategy illustrated a more balanced performance across the classifiers, with DT and SVM showing moderate improvements in precision and recall.

Lastly, the combination of all three feature sets in this second experiment underscored RF as the most robust classifier with an accuracy of 78.44% and an F1 of 79.09%. The SVM showed better balanced accuracy compared to previous setups, though it still struggled with recall.

4.6. Comparison with the State-of-the-Art

Table 14 showcases a comparative analysis of the performance of our work against previous state-of-the-art studies on the GasHisSDB dataset.

Table 14. Performance comparison of our work and the previous state-of-the-art works on the GasHisSDB dataset. * indicates that the proposed approach was trained on S-A and directly tested on S-B and S-C without fine-tuning.

Work	Split (%)	Model Details	A (%)		
			S-C	S-B	S-A
[2]	40/40/20	VGG16	96.12	96.47	95.90
	40/40/20	ResNet50	96.09	95.94	96.09
[46]	40/20/40	InceptionV3 trained from scratch	-	-	98.83
	40/20/40	InceptionV3 + ResNet50 (feature concatenation)	-	-	98.80
[12]	60/20/20	LGFFN	-	-	96.81
[47]	80/-/20	MCLNet based on ShuffleNetV2	96.28	97.95	97.85
[9]	40/20/40	Ensemble	97.72	98.68	99.20
Our	5-fold CVal	SVM with feature fusion	60.31*	85.82*	95.03
Our	5-fold CVal	RF with feature fusion	78.44*	89.56*	92.26

In Hu et al. [2], two models, VGG16 and ResNet50, were tested with a 40/40/20 split. VGG16 achieved accuracies of 96.12%, 96.47%, and 95.90% across S-C, S-B, and S-A, respectively. Similarly, ResNet50 showed comparable performance with 96.09%, 95.94%, and 96.09% in the same sub-databases.

 doi:10.20944/preprints202407.0478.v1

In the study by [46], an InceptionV3 model trained from scratch using a 40/20/40 split achieved a remarkable 98.83% accuracy in the S-A sub-database. Furthermore, combining InceptionV3 and ResNet50 through feature concatenation yielded a very close accuracy of 98.80%.

Li et al. [12] used a local-global feature fuse network (LGFFN) with a 60/20/20 split, achieving an accuracy of 96.81% in the S-A. This approach leverages the strengths of local and global features to improve classification performance. On the other hand, [47] employed MCLNet based on ShuffleNetV2 with an 80/-/20 split, reporting high accuracies of 96.28%, 97.95%, and 97.85% across S-C, S-B, and S-A, respectively.

The ensemble method adopted by [9] with a 40/20/40 split exhibited outstanding results, with accuracies of 97.72%, 98.68%, and 99.20% in S-C, S-B, and S-A, respectively. This ensemble approach amalgamates the strengths of multiple models, thereby achieving superior performance and robustness in classification tasks.

The main differences between the state-of-the-art and our work are that our evaluation followed a 5-fold Cval protocol; we used only HC features and features extracted from pre-trained, off-the-shelf CNNs to evaluate the extent to which non-specialized and non-tuned features can accomplish the binary classification task faced in this study. We tested on S-B and S-C by using only models trained on S-A to investigate the influence of the image resolution size in this scenario.

As can be seen, we reported two models: SVM with feature fusion and RF with feature fusion. The SVM model achieved accuracies of 60.31%, 85.82%, and 95.03% in S-C, S-B, and S-A, respectively. Similarly, the RF model showed accuracies of 78.44%, 89.56%, and 92.26% in the same categories. Although lower in S-B and S-C due to the training/testing strategy employed, compared to previous studies, these results highlight the potential of feature fusion techniques in improving classification performance, even without the need for a complex fine-tuning strategy that can be time-consuming and, above all, require a high amount of labelled data that can be complex in medical scenarios [48]. The 5-fold CV method ensures a more robust evaluation by repeatedly training and testing on different subsets of the data, thus providing a reliable estimate of the models' performance.

5. Discussion

This section analyzes the key aspects of our study. Specifically, Section 5.1 examines the relative performance and merits of HC versus deep features, whereas Section 5.2 discusses the outcomes of combining HC and deep features, analyzing how this fusion impacts the overall classification performance. Next, Section 5.3 evaluates the robustness and adaptability of our classification models across different magnifications of the considered dataset. Finally, Section 5.4 addresses the constraints and potential weaknesses of our study.

5.1. On the HC vs. Deep Features Comparison

The comparative analysis presented in Sections 4.2 and 4.3 reveals that, on the one hand, LBP consistently performs well among HC features, demonstrating robustness and reliability across different classifiers, even without exceptional performance. On the contrary, Haar features generally perform poorly, suggesting they are less suitable for this task.

On the other hand, deep features extracted from pre-trained CNNs, especially DenseNet-201 and DarkNet-53, consistently outperform HC features. This underscores the advantage of using deep learning models for feature extraction in complex tasks such as histopathological image classification, even without fine-tuning strategies.

In addition, the Random Forest classifier has shown strong performance with both HC and deep features, indicating its versatility and effectiveness in handling various feature types. More precisely, RF with features extracted from DenseNet-201 demonstrated their reliability for the task.

In summary, the detailed performance evaluation across various feature-classifier combinations provides valuable insights into the strengths and weaknesses of different approaches. The consistent superiority of deep features, particularly those from DenseNet-201 and EfficientNetB0, suggests a

 doi:10.20944/preprints202407.0478.v1

clear direction in this domain, emphasizing the integration of advanced deep learning techniques for enhanced classification accuracy and robustness. This is the main reason that motivated us to pick them along with LBP among the HC features to investigate feature fusion strategies.

5.2. On the Feature Fusion Performance

Across all feature fusion strategies, SVM consistently outperformed both DT and RF, demonstrating its superior ability to handle the diverse and complex feature sets derived from combining HC and deep features. The consistent high performance of SVM across various combinations suggests it is highly adaptable to different types of feature information.

RF, while generally strong, showed variability in performance depending on the feature combination, indicating that it might be more sensitive to the quality and type of features used. DT, on the other hand, consistently lagged behind SVM and RF, pointing to its relatively lower capacity to effectively utilize complex feature sets.

The combination of LBP, DenseNet-201, and EfficientNetB0 features, particularly when used with SVM, provides the most reliable and high-performing strategy for classifying histopathological images. This fusion strategy leverages the strengths of both HC and deep features, resulting in a robust classification framework.

5.3. On the Cross-Magnification Performance

Overall, the experiments reveal that the RF classifier consistently outperforms DT and SVM across various feature fusion strategies and test set dimensions. The combination of DenseNet-201 and EfficientNetB0 generally provides the most reliable feature set, enhancing classifier performance and demonstrating the feasibility of using features provided by HC methods or pre-trained CNNs in the task of histopathological image classification.

The comprehensive analysis demonstrates the efficacy of feature fusion and the importance of choosing robust classifiers to achieve high accuracy and reliability in medical image classification tasks.

5.4. Limitations

This study, while comprehensive in its approach to evaluating the performance of shallow learning classifiers on histopathological image classification, presents several limitations that must be acknowledged.

First, the reliance on pre-trained CNNs for feature extraction, without any fine-tuning specific to the dataset at hand, may limit the potential performance of the classifiers. Fine-tuning these networks could potentially yield features more tailored to the specific characteristics of the histopathological images, thereby improving classification accuracy.

Second, the experiments were conducted using a single dataset, GasHisSDB, which might limit the generalizability of the findings. The performance measures observed might vary significantly when applied to other histopathological datasets with different image characteristics, variations in staining procedures, or differing disease profiles. A broader validation across multiple datasets would provide more robust evidence of the classifiers' effectiveness.

Third, the study employed a specific image resolution sub-database (160×160 for training and testing, 120×120 and 80×80 for testing). The impact of image resolution on classifier performance was not extensively explored, and it is possible that different resolutions could influence the feature extraction and classification processes.

Additionally, the study did not incorporate image augmentation techniques, which are commonly used in image classification tasks to improve model generalization by artificially increasing the size and variability of the training dataset. The absence of augmentation may result in overfitting, particularly given the limited data available for training.

 doi:10.20944/preprints202407.0478.v1

In summary, while this work provides valuable insights into the use of shallow learning classifiers for histopathological image classification with features provided by non-fine-tuned methods, the limitations discussed here suggest avenues for further research to enhance and validate the findings.

6. Conclusions

This study comprehensively evaluates shallow learning classifiers for histopathological image classification using HC and deep features.

The comparative analysis of HC versus deep features demonstrates the clear superiority of deep features, particularly those extracted from pre-trained CNNs such as DenseNet-201 and EfficientNetB0. These features consistently outperform HC features, highlighting the advanced feature extraction capability of DL models in complex image classification tasks. Among HC features, LBP shows robust performance, while Haar features are less effective.

Our exploration of feature fusion techniques shows that combining features can significantly enhance classification performance. The SVM classifier, in particular, excels in handling diverse and complex feature sets, outperforming both DT and RF classifiers across various combinations. The fusion of LBP, DenseNet-201, and EfficientNetB0 features emerges as the most reliable strategy, leveraging the strengths of both HC and deep features.

In addition, our cross-magnification experiments underscore the robustness of RF classifiers, which consistently perform well across different image resolutions. The combination of features from DenseNet-201 and EfficientNetB0 proves effective in maintaining high classification accuracy, demonstrating the feasibility of utilizing features from pre-trained CNNs and HC methods.

While this study advances our understanding of shallow learning classifiers in histopathological image classification, our results open the field for several future works. For instance, with feature importance and explainability techniques, we plan to investigate the most effective features for the classifiers' final prediction to simplify the workflow with feature selection strategies. Moreover, we aim to integrate further DL methods like Vision Transformer and leverage them as feature extractors in this context.

Author Contributions: Conceptualization, A.L. and C.D.R.; methodology, M.U., A.L. and C.D.R.; investigation, M.U., A.L. and C.D.R.; software, M.U. and A.L.; writing—original draft, M.U. and A.L.; writing—review and editing, M.U., A.L., and C.D.R.; supervision, A.L. and C.D.R.. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: All the material used and developed for this work is available at the following GitHub repository: https://github.com/MurkoZawa/HistopathologyClassification (accessed on 2 July 2024).

Acknowledgments: We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5—Call for tender No. 3277 published on 30 December 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union—NextGenerationEU. Project Code ECS0000038—Project Title eINS Ecosystem of Innovation for Next Generation Sardinia—CUP F53C22000430001-Grant Assignment Decree No. 1056 adopted on 23 June 2022 by the Italian Ministry of University and Research (MUR) and by the project DEMON "Detect and Evaluate Manipulation of ONline information" funded by MIUR under the PRIN 2022 grant 2022BAXSPY (CUP F53D23004270006, NextGenerationEU).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
DL	Deep Learning
HC	Handcrafted
CNN	Convolutional Neural Network
GC	Gastric Cancer

 doi:10.20944/preprints202407.0478.v1

EGC	early-stage gastric cancer
AGC	advanced gastric cancer
CV	Computer Vision
WSI	whole slide image
LGFFN	lightweight gated fully-fused network
GHI	gated hybrid input
CH	Chebyshev Moments
LM	Second-order Legendre Moments
ZM	Zernike Moments
HAR	Rotation-Invariant Haralick
LBP	Local Binary Pattern
Hist	Histogram
AC	Autocorrelogram
Haar	Haar-like
DT	Decision Tree
kNN	k-Nearest Neighbor
SVM	Support Vector Machine
RF	Random Forest
TN	True Negatives
FP	False Positives
FN	False Negatives
TP	True Positives
A	Accuracy
P	Precision
R	Recall
S	Specificity
F	F1-score
MCC	Matthews Correlation Coefficient
BACC	Balanced Accuracy
Cval	cross-validation

References

1. Ilic, M.; Ilic, I. Epidemiology of stomach cancer. *World journal of gastroenterology* **2022**, *28*, 1187.
2. Hu, W.; Li, C.; Li, X.; Rahaman, M.M.; Ma, J.; Zhang, Y.; Chen, H.; Liu, W.; Sun, C.; Yao, Y.; Sun, H.; Grzegorzek, M. GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Comput. Biol. Medicine* **2022**, *142*, 105207. doi:10.1016/J.COMPBIOMED.2021.105207.
3. Zhang, K.; Wang, H.; Cheng, Y.; Liu, H.; Gong, Q.; Zeng, Q.; Zhang, T.; Wei, G.; Wei, Z.; Chen, D. Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images. *Scientific Reports* **2024**, *14*, 7847.
4. Hirasawa, T.; Aoyama, K.; Tanimoto, T.; Ishihara, S.; Shichijo, S.; Ozawa, T.; Ohnishi, T.; Fujishiro, M.; Matsuo, K.; Fujisaki, J.; others. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* **2018**, *21*, 653–660.
5. Yoon, H.J.; Kim, S.; Kim, J.H.; Keum, J.S.; Oh, S.I.; Jo, J.; Chun, J.; Youn, Y.H.; Park, H.; Kwon, I.G.; others. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. *Journal of clinical medicine* **2019**, *8*, 1310.
6. Zhao, Y.; Hu, B.; Wang, Y.; Yin, X.; Jiang, Y.; Zhu, X. Identification of gastric cancer with convolutional neural networks: a systematic review. *Multim. Tools Appl.* **2022**, *81*, 11717–11736. doi:10.1007/S11042-022-12258-8.
7. Xie, K.; Peng, J. Deep learning-based gastric cancer diagnosis and clinical management. *Journal of Radiation Research and Applied Sciences* **2023**, *16*, 100602.
8. Hu, W.; Chen, H.; Liu, W.; Li, X.; Sun, H.; Huang, X.; Grzegorzek, M.; Li, C. A comparative study of gastric histopathology sub-size image classification: From linear regression to visual transformer. *Frontiers in Medicine* **2022**, *9*, 1072109.
9. 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.

10. Cao, R.; Tang, L.; Fang, M.; Zhong, L.; Wang, S.; Gong, L.; Li, J.; Dong, D.; Tian, J. Artificial intelligence in gastric cancer: applications and challenges. *Gastroenterology Report* **2022**, *10*, goac064.
11. Hu, W.; Li, C.; Rahaman, M.M.; Chen, H.; Liu, W.; Yao, Y.; Sun, H.; Grzegorzek, M.; Li, X. EBHI: A new Enteroscope Biopsy Histopathological H&E Image Dataset for image classification evaluation. *Physica Medica* **2023**, *107*, 102534.
12. Li, S.; Liu, W. LGFFN-GHI: A Local-Global Feature Fuse Network for Gastric Histopathological Image Classification. *Journal of Computer and Communications* **2022**, *10*, 91–106.
13. Putzu, L.; Loddo, A.; Ruberto, C.D. Invariant Moments, Textural and Deep Features for Diagnostic MR and CT Image Retrieval. Computer Analysis of Images and Patterns - 19th International Conference, CAIP 2021, Virtual Event, September 28-30, 2021, Proceedings, Part I; Tsapatsoulis, N.; Panayides, A.; Theocharides, T.; Lanitis, A.; Pattichis, C.S.; Vento, M., Eds. Springer, 2021, Vol. 13052, *Lecture Notes in Computer Science*, pp. 287–297. doi:10.1007/978-3-030-89128-2_28.
14. Ruberto, C.D.; Loddo, A.; Putzu, L. On The Potential of Image Moments for Medical Diagnosis. *J. Imaging* **2023**, *9*, 70. doi:10.3390/JIMAGING9030070.
15. Mukundan, R.; Ong, S.H.; Lee, P.A. Image analysis by Tchebichef moments. *IEEE Trans. Image Process.* **2001**, *10*, 1357–1364. doi:10.1109/83.941859.
16. Ruberto, C.D.; Putzu, L.; Rodriguez, G. Fast and accurate computation of orthogonal moments for texture analysis. *Pattern Recognit.* **2018**, *83*, 498–510. doi:10.1016/J.PATCOG.2018.06.012.
17. Teh, C.; Chin, R.T. On Image Analysis by the Methods of Moments. *IEEE Trans. Pattern Anal. Mach. Intell.* **1988**, *10*, 496–513. doi:10.1109/34.3913.
18. Teague, M.R. Image analysis via the general theory of moments. *Josa* **1980**, *70*, 920–930.
19. Wee, C.; Raveendran, P. On the computational aspects of Zernike moments. *Image Vis. Comput.* **2007**, *25*, 967–980. doi:10.1016/J.IMAVIS.2006.07.010.
20. Mirjalili, F.; Hardeberg, J.Y. On the Quantification of Visual Texture Complexity. *J. Imaging* **2022**, *8*, 248. doi:10.3390/JIMAGING8090248.
21. Putzu, L.; Ruberto, C.D. Rotation Invariant Co-occurrence Matrix Features. Image Analysis and Processing - ICIAP 2017 - 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I; Battiato, S.; Gallo, G.; Schettini, R.; Stanco, F., Eds. Springer, 2017, Vol. 10484, *Lecture Notes in Computer Science*, pp. 391–401. doi:10.1007/978-3-319-68560-1_35.
22. He, D.C.; Wang, L. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing* **1990**, *28*, 509–512.
23. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. doi:10.1109/TPAMI.2002.1017623.
24. van de Weijer, J.; Schmid, C. Coloring Local Feature Extraction. Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part II; Leonardis, A.; Bischof, H.; Pinz, A., Eds. Springer, 2006, Vol. 3952, *Lecture Notes in Computer Science*, pp. 334–348. doi:10.1007/11744047_26.
25. Huang, J.; Kumar, R.; Mitra, M.; Zhu, W.; Zabih, R. Image Indexing Using Color Correlograms. 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico. IEEE Computer Society, 1997, pp. 762–768. doi:10.1109/CVPR.1997.609412.
26. Viola, P.A.; Jones, M.J. Rapid Object Detection using a Boosted Cascade of Simple Features. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA. IEEE Computer Society, 2001, pp. 511–518. doi:10.1109/CVPR.2001.990517.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. doi:10.1145/3065386.
28. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 6517–6525. doi:10.1109/CVPR.2017.690.
29. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *CoRR* **2018**, *abs/1804.02767*, [1804.02767].
30. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

31. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA; Chaudhuri, K.; Salakhutdinov, R., Eds. PMLR, 2019, Vol. 97, *Proceedings of Machine Learning Research*, pp. 6105–6114.
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.
33. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA; Singh, S.; Markovitch, S., Eds. AAAI Press, 2017, pp. 4278–4284. doi:10.1609/AAAI.V31I1.11231.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; Bengio, Y.; LeCun, Y., Eds., 2015.
36. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 1800–1807. doi:10.1109/CVPR.2017.195.
37. Bodapati, J.D.; Veeranjaneyulu, N. Feature Extraction and Classification UsingDeep Convolutional Neural Networks. *J. Cyber Secur. Mobil.* **2019**, *8*, 261–276. doi:10.13052/JCSM2245-1439.825.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
39. Putzu, L.; Piras, L.; Giacinto, G. Convolutional neural networks for relevance feedback in content based image retrieval. *Multim. Tools Appl.* **2020**, *79*, 26995–27021. doi:10.1007/S11042-020-09292-9.
40. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020, pp. 8681–8691. doi:10.1109/CVPR42600.2020.00871.
41. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015; Bach, F.R.; Blei, D.M., Eds. JMLR.org, 2015, Vol. 37, *JMLR Workshop and Conference Proceedings*, pp. 448–456.
42. Quinlan, J.R. Learning efficient classification procedures and their application to chess end games. In *Machine learning*; Springer, 1983; pp. 463–482.
43. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.
44. Lin, Y.; Lv, F.; Zhu, S.; Yang, M.; Cour, T.; Yu, K.; Cao, L.; Huang, T.S. Large-scale image classification: Fast feature extraction and SVM training. The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011. IEEE Computer Society, 2011, pp. 1689–1696.
45. Breiman, L. Random Forests. *Machine Learning* **2001**, *4*, 5–32.
46. Springenberg, M.; Frommholz, A.; Wenzel, M.; Weicken, E.; Ma, J.; Strodthoff, N. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. *Medical Image Anal.* **2023**, *87*, 102809. doi:10.1016/J.MEDIA.2023.102809.
47. 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. doi:10.1016/J.BSPC.2022.104319.
48. Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys* **2023**, *55*, 1–40.