Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

LEAKAGE-FREE HYBRID DEEP FEATURE FUSION AND CATBOOST FOR GASTRIC CANCER DETECTION IN HISTOPATHOLOGICAL IMAGES

Muhammad Hamza Afzal[*1], Qamar Farooq[2], Qamar Ayyub[3], Asad Ullah Gill[4], Haroon Noor[5]

[*1,2,3,4,5]*The Superior University Faisalabad Campus*

[*1]ihamzaafzal@gmail.com, [2] qamar farooq, [3]qamarayyub@gmail.com,
[4]asadulahgill17@gmail.com, [5]haroonkhan550@gmail.com

DOI: https://doi.org/10.5281/zenodo.20303291

Keywords
Gastric cancer; histopathology image classification; GasHisSDB; deep feature fusion; CatBoost; explainable artificial intelligence; LIME

Article History
Received: 22 March 2026
Accepted: 01 May 2026
Published: 20 May 2026

Corresponding Author: *
Muhammad Hamza Afzal

Abstract

Gastric cancer is still a significant global health problem and histopathological examination plays a crucial role in the definitive diagnosis. But, manual microscopic examination is time consuming and may be inconsistent between different observers, which has led to the desire for reliable computer-aided diagnostic systems. In this work, a leakage-free hybrid scheme for binary classification of gastric histopathology images in "abnormal" class and "normal" class based on GasHisSDB160 data set is proposed. To ensure no overlap between the data partitions, a group-aware split was applied to partition 33,284 image patches into training, validation, and independent test sets. The framework is a mix of custom Convolutional Network (CNN), ResNet50V2 and MobileNetV2. The two strategies that were tested were an end-to-end deep learning classifier and a CatBoost classifier trained on PCA-reduced fused features where no leakage was observed. The deep learning model obtained a test accuracy of 97.94%, macro F1-score of 97.85% and ROC-AUC of 99.78%. The proposed CatBoost model achieved maximum test accuracy of 98.28%, macro F1-score: 98.21% and Matthews correlation coefficient (MCC): 96.41% with 86 test images misclassified out of the total of 4,993 images. LIME visual explanations also helped to explain decisions made by the model. The findings show that leakage-aware deep feature fusion with CatBoost is a competitive and explainable method in the support of gastric cancer screening in digital pathology.

1. INTRODUCTION

Gastric cancer is one of the most clinically relevant cancers of the world. According to the latest data on cancer cases globally, there were about 969,000 new cases of stomach cancer in 2022, with 660,000 deaths, making it one of the top 10 most common cancers and the top 10 most deadly cancers.[1] It is important, therefore, to detect these early and accurately so that treatment plans can be made in a timely manner and help the patient to get better. Histopathological examination using hematoxylin and eosin (H&E) stained tissue still serves as the diagnostic standard, but is time consuming, subjective and relies on the availability of a trained pathologist for the diagnosis. Digital pathology and machine learning provide an alternative approach to aiding pathologists to scan a vast number of image patches and noting down the suspicious ones. While convolutional neural networks (CNNs) can be used to learn discriminative visual patterns from histopathology images [1], single-network models may not be able to learn all tissue-level and

texture-level patterns needed for robust decision making. Models like ResNet50V2 and MobileNetV2 [2], [3], [4] are good pretrained representations that can be used to build transfer learning models, and custom CNN layers can learn low- and mid-level cues specific to the dataset. This merging of representations can help the robustness of the classification since different kinds of visual evidence are learned by each branch.

Despite the promising performance of deep learning in histopathology, research validity depends strongly on the experimental protocol. Data leakage can occur when image patches from the same patient, slide, or source distribution are shared between training and testing. Leakage can inflate performance and reduce clinical relevance. This study therefore emphasizes a leakage-free workflow in which training, validation, and test partitions are separated before feature extraction, and PCA is fitted only on training features. The final independent test set is used only once for reporting.

The following are the main contributions of this study. First, a leakage-free hybrid framework for binary gastric histopathology classification is developed by using GasHisSDB160. Second, complementary visual representations are taken advantage of by fused deep features from a custom CNN, ResNet50V2, and MobileNetV2. Third, PCA is applied without data leakage to obtain a second feature vector of 224 dimensions from the fused 3,456-dimensional feature vector to pass on to CatBoost classification. Fourth, clinically relevant metrics such as sensitivity, specificity, false-negative rate, false-positive rate, Matthews correlation coefficient, ROC-AUC, and average precision are reported. Fifth, the local interpretability is illustrated through LIME explanations for the prediction of abnormal and normal tissue classes.

2. Related Work

The GasHisSDB dataset was introduced to address the shortage of publicly available gastric histopathology image datasets. It contains 245,196 sub-size image patches derived from gastric H&E histopathology, divided into normal and abnormal categories across different patch sizes. The original benchmark evaluated classical machine learning, CNN-based methods, and transformer-based models, reporting a best deep learning accuracy of 96.47%. This established GasHisSDB as an important benchmark for computer-aided gastric cancer diagnosis. [5]

In the following studies, ensemble learning and feature fusion have been used to boost the performance. Yong et al. (2023) [1] have used deep ensemble learning and reported high accuracy on GasHisSDB including 99.20% accuracy for the subset (160 x 160 pixels). Mudavadkar et al. (2024) [6] also showed that ensemble models can achieve an average accuracy exceeding 99% for all subset patch sizes of GasHisSDB, showing a good performance for 160 x 160 patch size subset. Loddo et al. (2024) [7] also investigated feature fusion techniques and evaluated different classifiers for gastric image classification, demonstrating that the selection of the appropriate representation and classifier significantly impacts the classification accuracy.

Hybrid deep feature extraction with shallow or gradient-boosting classifiers is attractive when the feature representation is strong but end-to-end classification may overfit. [8] CatBoost is a gradient boosting algorithm that uses ordered boosting to reduce prediction shift and is robust across many datasets. [9] Although CatBoost was originally designed with special handling for categorical features, it is also effective on compact numerical representations such as PCA-transformed deep features. LIME, meanwhile, provides local explanations by perturbing input regions and fitting interpretable local surrogate models. [10] In medical imaging, such visual explanations are important because diagnostic accuracy alone is insufficient for clinical trust.

Compared with prior studies, the novelty of the present work is not merely high accuracy; rather, the emphasis is on a reproducible, leakage-aware fusion pipeline in which deep learning, PCA, CatBoost, and explainability are combined while preserving strict separation of training, validation, and test data. This is important because reported performance values in digital

pathology can be misleading when image-level splitting allows hidden overlap between partitions.

3. Methodology

3.1 Dataset

The study used the GasHisSDB160 sub-database [5], consisting of 33,284 H&E-stained gastric histopathology patches. The binary classification task distinguishes abnormal gastric cancer patches from normal gastric tissue patches. The dataset contained 13,124 abnormal images and 20,160 normal images. All images were resized to 224 x 224 pixels before model training to satisfy the input requirement of the transfer learning branches.

Table 1. Overall GasHisSDB160 class distribution

Class	Number of images	Percentage
Abnormal	13,124	39.43%
Normal	20,160	60.57%
Total	33,284	100.00%

3.2 Leakage-Free Data Partitioning

Training was performed on a split with group awareness. Group Id's were allocated based on the image name in the file. The split resulted in about 70% for training, 15% for validation and 15% for final testing. It was split into 23,297 training images, 4,994 validation images and 4,993 test images. The leakage check indicated that there were no overlapping groups between training and validation, training and test, and validation and test data sets.

Table 2. Leakage-free split size summary

Split	Images	Groups	Percentage
Training	23,297	23,297	69.99%
Validation	4,994	4,994	15.00%
Testing	4,993	4,993	15.00%

Table 3. Class distribution across partitions

Class	Training	Validation	Testing
Normal	14,141	3,025	2,994
Abnormal	9,156	1,969	1,999

Table 4. Group-level leakage check

Comparison	Overlapping groups
Train vs Validation	0
Train vs Test	0
Validation vs Test	0

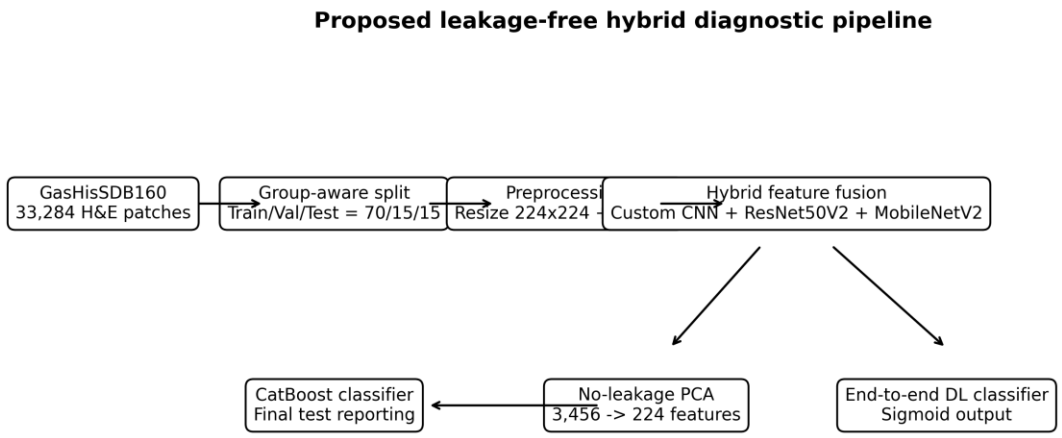

Figure 1. Proposed leakage-free hybrid diagnostic pipeline.

3.3 Preprocessing and Augmentation

All the images were scaled by dividing the intensity of the pixels by 255. Random rotation, zooming, and horizontal flipping were used to augment the training images. Validation and test images were not augmented and were not shuffled, thus providing reproducible evaluation and proper alignment between the true label and prediction. This design was only for augmenting the training data to introduce varied training data, not to change the data used for validation and final testing.

3.4 Hybrid Deep Feature Fusion

The proposed neural network has three feature extraction arms. The first branch consists of a custom CNN composed of three blocks of convmax-pool + GAP with 32, 64 and 128 filters, respectively. The second branch is a ResNet50V2 model [3], [2] that is pre-trained on ImageNet and fine-tuned in the last layers while the first layers are frozen. The third branch is the MobileNetV2 [4], pretrained on ImageNet, but without the classification head. The pretrained branches have been globally averaged. The three branches were joined together to create a 3,456 dimensional fused feature vector. The end to end deep learning model consisted of fully connected layers with neurons of 512, 256, 128, 64, and 32 and in between them dropout layers of 0.50, 0.40, 0.30, 0.25, and 0.20, respectively. A sigmoid output neuron was used for binary classification. The model was optimized with binary cross-entropy loss and adam optimizer with learning rate of 1e-4. Both early stopping and learning rate reduction used to monitor validation accuracy and validation loss, respectively. The final network used 27,860,609 total parameters, of which 18,705,217 were trainable, and 9,155,392 were not.

Table 5. Summary of the hybrid deep learning architecture

Component	Configuration	Purpose
Custom CNN	Three convolutional blocks	Dataset-specific local texture features
ResNet50V2	ImageNet-pretrained residual network	Deep semantic and residual features
MobileNetV2	ImageNet-pretrained lightweight network	Efficient inverted-residual features
Fusion layer	Concatenation	3,456-dimensional fused representation
Dense classifier	512-256-128-64-32 + dropout	End-to-end sigmoid classification

3.5 CatBoost on No-Leakage PCA Features

In the second strategy the feature extractor was the fused feature layer. The feature extraction was performed on the original training, validation, and test generators. There were 23297 samples and 3456 features in the training feature matrix, 4994 samples and 3456 features in the validation feature matrix, and 4993 samples and 3456 features in the test feature matrix. Only the training features were fitted with PCA [11] and this was used to transform validation and test features. This ensured that information from validation or test partitions did not have an impact on the dimensionality reduction. The maximum number of PCA components was set to 224, resulting in training, validation, and test matrices of 23,297 x 224, 4,994 x 224, and 4,993 x 224, respectively.

CatBoost [9] was trained on the PCA-transformed training features with validation-based overfitting detection. The best validation accuracy was obtained at iteration 327, after which the model was shrunk to the first 328 iterations. Final predictions were made only on the independent test set.

Table 6. CatBoost feature-reduction and training details

Parameter	Value
Raw fused feature dimension	3,456
PCA fitting data	Training set only
PCA components retained	224
CatBoost training samples	23,297
CatBoost validation samples	4,994
CatBoost final test samples	4,993
Best validation iteration	327
Final CatBoost trees retained	328

3.6 Evaluation Metrics

For the evaluation of the models, accuracy, precision, recall, F1-score, ROC-AUC, average precision, specificity, negative predictive value, balanced accuracy, false-positive rate, false-negative rate, and Matthews correlation coefficient (MCC) were utilized. For clinically oriented metrics, abnormal gastric tissue was considered the positive class. The main formulas were: Accuracy = (TP + TN)/(TP + TN + FP + FN), Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP), Precision = TP/(TP + FP), F1 = 2 x Precision x Recall/(Precision + Recall), and MCC = (TP x TN - FP x FN)/sqrt((TP + FP)(TP + FN)(TN + FP)(TN + FN)).

4. Results

4.1 Training and Validation Performance

The end-to-end deep learning model gradually improved during the training process. The

validation accuracy was improved from 80.36% at the first epoch to 98.14% at the restored best checkpoint. The model validation loss was 0.0607, the validation accuracy was 98.14%, the precision was 98.54% and the recall was 98.38% for validation data. These values reflect the fact that the feature network obtained by fusing the features learned stable discriminative patterns prior to the ultimate testing.

Table 7. Validation performance of the deep learning model

Metric	Value
Validation loss	0.060676
Validation accuracy	0.981378
Validation precision	0.985430
Validation recall	0.983802

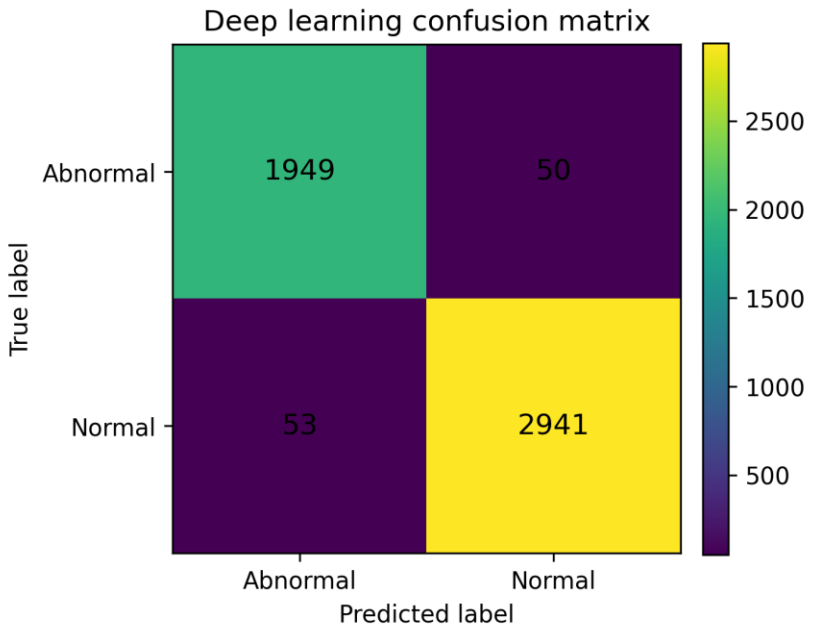

Figure 2. Confusion matrix of the end-to-end deep learning model on the independent test set.

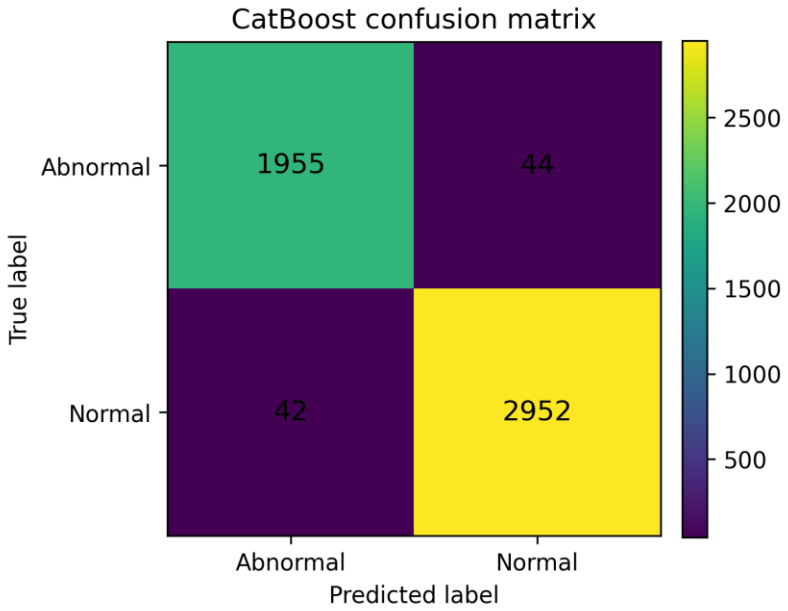

Figure 3. Confusion matrix of the CatBoost model on the independent test set.

4.2 Independent Test Performance

There were 4,993 images in the independent test set. In the end to end deep learning model, 4,890 test images were classified correctly while 103 errors were made. The CatBoost model was able to classify 4,907 test images correctly with 86 errors. As a result, CatBoost decreased the quantity of test errors by 16.50% in comparison with the end-to-end deep learning classifier. This improvement was also observed in terms of accuracy, macro recall, macro F1-score and MCC.

Table 8. Main independent test performance comparison

Model	Accuracy	Prec. Macro	Recall Macro	F1 Macro	Prec. Weighted	Recall Weighted	F1 Weighted	ROC-AUC	Avg. Precision
Deep learning fusion model	0.979371	0.978405	0.978643	0.978523	0.979377	0.979371	0.979374	0.997781	0.998232
PCA + CatBoost on fused features	0.982776	0.982141	0.981980	0.982061	0.982773	0.982776	0.982774	0.997519	0.997967

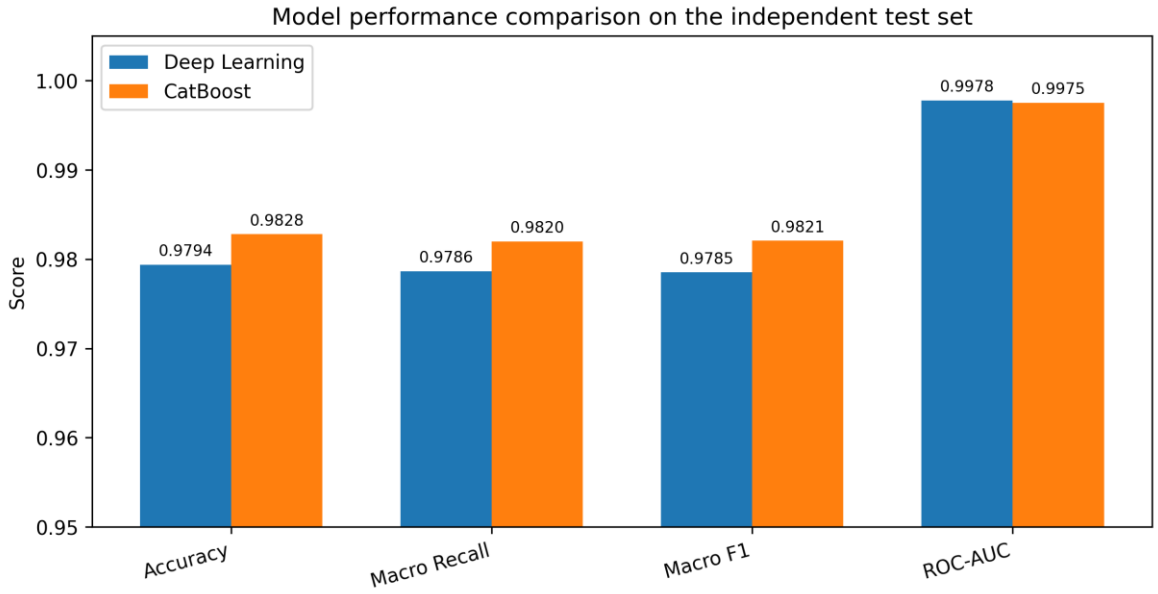

Figure 4. Comparison of major test metrics for the deep learning and CatBoost models.

Table 9. Confusion-matrix counts on the independent test set

Model	True Abnormal	False Normal	False Abnormal	True Normal	Total errors
Deep learning fusion model	1,949	50	53	2,941	103
PCA + CatBoost on fused features	1,955	44	42	2,952	86

Table 10. Clinically oriented diagnostic metrics using abnormal tissue as the positive class

Model	Sensitivity	Specificity	PPV	NPV	Balanced Accuracy	MCC	FPR	FNR
Deep learning fusion model	97.50%	98.23%	97.35%	98.33%	97.86%	95.70%	1.77%	2.50%
PCA + CatBoost on fused features	97.80%	98.60%	97.90%	98.53%	98.20%	96.41%	1.40%	2.20%

4.3 Per-Class Results

CatBoost model resulted in a balanced performance for both the diagnostic categories. For abnormal images, precision was 97.90%, recall was 97.80%, and F1-score was 97.85%. For normal images, precision was 98.53%, recall was 98.60%, and F1-score was 98.56%. The small difference between the abnormal and normal indicates that the model was not simply exploiting the majority normal class, rather the model maintained high recognition rate for the clinically more important abnormal class.

Table 11. CatBoost per-class classification report

Class	Precision	Recall	F1-score	Support
Abnormal	0.978968	0.977989	0.978478	1,999
Normal	0.985314	0.985972	0.985643	2,994
Macro average	0.982141	0.981980	0.982061	4,993
Weighted average	0.982773	0.982776	0.982774	4,993

4.4 Model Explainability

Representative test images of abnormal and normal images were subjected to LIME application [10]. The model gave the probability of the abnormal class for the selected abnormal example, which was close to 1.0000. In the chosen normal sample, the model predicted the normal class, the probability of normal class was 0.9816, and the probability of abnormal class was 0.0184. The superpixels highlighted on the image are the ones that most strongly contributed to the local prediction. These visual explanations can be beneficial in medical imaging analysis as they enable researchers and medical professionals to determine if the model is focusing on the tissue structures or on irrelevant background patterns.

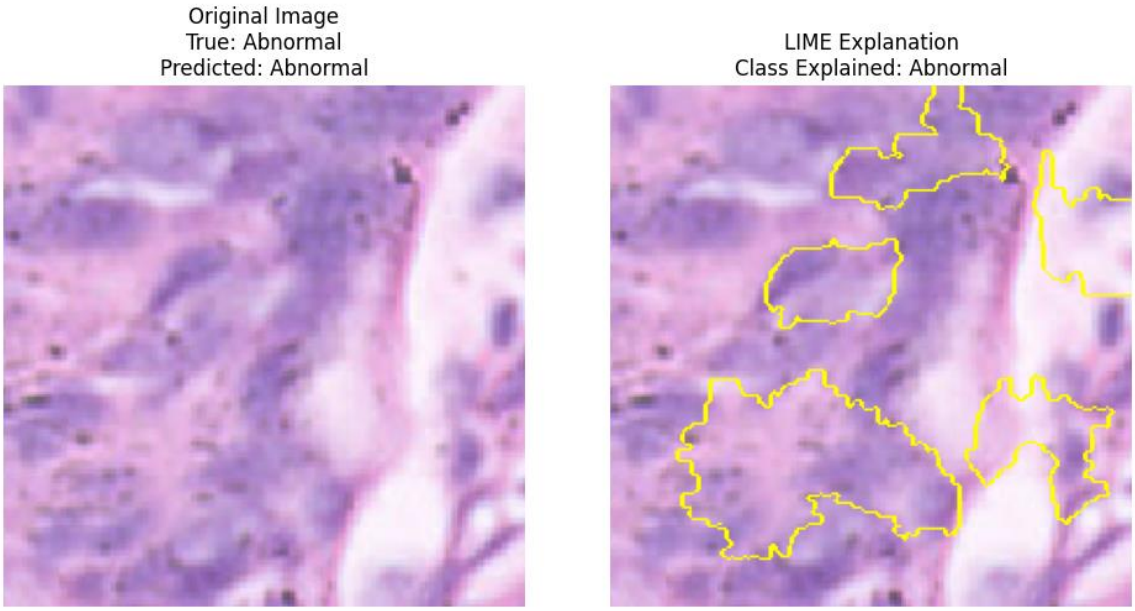

Figure 5. LIME explanation for a correctly classified abnormal gastric histopathology image.

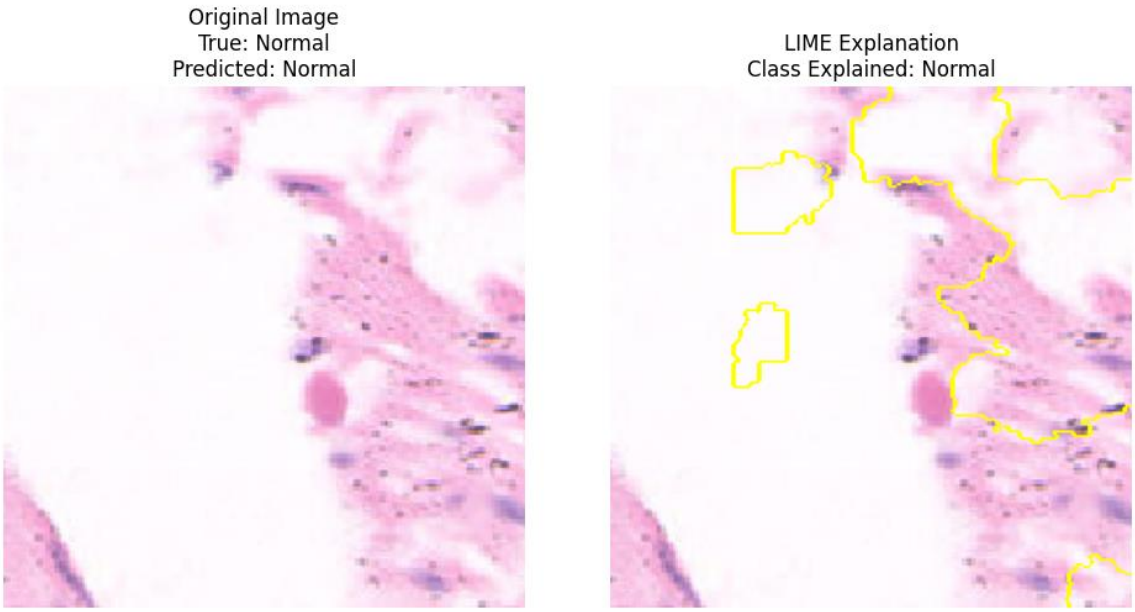

Figure 6. LIME explanation for a correctly classified normal gastric histopathology image.

5. Discussion

The results demonstrate that a hybrid deep feature fusion framework can produce strong diagnostic performance on GasHisSDB160 [5]. The end-to-end deep learning model already achieved high test performance, with 97.94% accuracy and 99.78% ROC-AUC. However, using the fused deep representation as input to PCA and CatBoost [11], [9] further improved accuracy to 98.28% and macro F1-score to 98.21%. The CatBoost model also reduced false abnormal predictions from 53 to 42 and false normal predictions from 50 to 44. Since false normal predictions are clinically important because abnormal tissue may be missed, the reduction from 50 to 44 is meaningful.

The performance gain suggests that the fused feature vector contains useful discriminative information that can be exploited by a non-neural classifier. Deep networks are optimized end to end, but the final dense classifier may not always use the fused feature space optimally, particularly when the dataset is moderately sized and class distribution is imbalanced. PCA reduces redundancy in the feature vector, while CatBoost learns non-linear decision boundaries over compact feature components. This combination appears to improve generalization without retraining the full CNN stack.

A key result is that macro and weighted metrics are close to each other. The accuracy may be overrepresenting performance as there are more normal than abnormal images in the dataset. Macro F1 score: This is a F1 score in which both abnormal class and normal class are equally important. The macro F1 score of 98.21 % and balanced accuracy of 98.20 % indicates that the performance is high not only for majority class, but also for minority class. The MCC is 96.41% which is another indicator of good binary classification in the presence of class imbalance.

Compared with the original GasHisSDB benchmark [5], the proposed CatBoost model exceeds the reported best deep learning accuracy of 96.47%. More recent ensemble studies have reported around 99% accuracy or 99.20% accuracy on the 160 x 160 pixel subset [6], [1], so the present framework should be interpreted as competitive rather than universally state-of-the-art. Direct comparison is difficult because studies may differ in split strategy, preprocessing, hyperparameter search, augmentation, and leakage control. The strength of the present work is its clear leakage-free design, independent test

reporting, detailed diagnostic metrics, and explainability using LIME.

The model was of great relevance in computer aided pathology. High sensitivity and specificity for a screening system may aid in prioritizing suspicious areas for pathologists to review, decreasing workload, and enabling quality control. But the model shouldn't be used as a standalone diagnostic replacement. Outside the context of the clinic, external validation of multi-center whole-slide images, prospective evaluation, robustness testing across different scanners and staining protocols and integration in workflows where the pathologist is in the loop are required.

Table 12. Comparison with selected GasHisSDB-related studies

Study	Approach	Reported/observed result	Relevance
[5]	GasHisSDB benchmark using classical ML, CNNs, and transformer models	Best deep learning accuracy: 96.47%	Established GasHisSDB as a benchmark
[1]	Deep ensemble learning on GasHisSDB	Highest reported accuracy on 160 x 160 subset: 99.20%	Strong ensemble benchmark
[7]	Ensemble digital pathology framework	Ensemble accuracy around 99% for 160 x 160 subset	Demonstrated value of ensemble learning
[7]	Comparative analysis and feature fusion strategies	Showed feature/classifier choices affect performance	Supports fusion-based evaluation
Proposed study	Leakage-free CNN + ResNet50V2 + MobileNetV2 features with PCA + CatBoost	Accuracy: 98.28%; macro F1: 98.21%; MCC: 96.41%	Competitive, leakage-aware, explainable hybrid pipeline

6. Conclusion

This study presented a leakage-free hybrid deep feature fusion framework for gastric cancer detection in histopathology image patches. The proposed method combined a custom CNN, ResNet50V2, and MobileNetV2 to generate rich fused features, followed by no-leakage PCA and CatBoost classification. On an independent test set of 4,993 GasHisSDB160 images, the CatBoost model achieved 98.28% accuracy, 98.21% macro F1-score, 99.75% ROC-AUC, and 96.41% MCC. It outperformed the end-to-end deep learning classifier in accuracy, macro F1-score, balanced accuracy, and total error count. LIME explanations provided local visual evidence for model decisions. The results support the use of leakage-aware hybrid feature fusion and gradient boosting as a strong approach for computer-aided gastric histopathology screening. Future work should include multi-center external validation, whole-slide-level aggregation, Grad-CAM comparison, prospective testing, and more extensive ablation studies across individual feature branches.

Data Availability Statement

The data used in this study are from the publicly available GasHisSDB dataset [5]. The experimental notebook and derived result files can be made available by the corresponding author upon reasonable request, subject to institutional and journal requirements.

REFERENCES

[1] M. P. Yong *et al.*, "Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning," *Diagnostics*, vol. 13, no. 10, p. 1793, May 2023, doi: 10.3390/diagnostics13101793.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, *arXiv*. doi: 10.48550/ARXIV.1512.03385.

[3] S. Zhang, X. Shen, Z. Lin, R. Mech, J. P. Costeira, and J. M. F. Moura, "Learning to Understand Image Blur," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 6586–6595. doi: 10.1109/CVPR.2018.00689.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[5] "W. Hu, C. Li, X. Li, M. M. Rahaman, J. Ma, Y. Zhang, H. Chen, W. Liu, C. Sun, Y. Yao, H. Sun, and M. Grzegorzek, 'GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer,' Computer Methods and Programs in Biomedicine, vol. 216, Art. no. 105207, 2022,".

[6] G. R. Mudavadkar *et al.*, "Gastric Cancer Detection with Ensemble Learning on Digital Pathology: Use Case of Gastric Cancer on GasHisSDB Dataset," *Diagnostics*, vol. 14, no. 16, p. 1746, Aug. 2024, doi: 10.3390/diagnostics14161746.

[7] A. Loddo, M. Usai, and C. Di Ruberto, "Gastric Cancer Image Classification: A Comparative Analysis and Feature Fusion Strategies," *J. Imaging*, vol. 10, no. 8, p. 195, Aug. 2024, doi: 10.3390/jimaging10080195.

[8] D. Khayatian, A. Maleki, H. Nasiri, and M. Dorrigiv, "Histopathology image analysis for gastric cancer detection: a hybrid deep learning and catboost approach," *Multimed. Tools Appl.*, vol. 84, no. 19, pp. 21777–21803, Aug. 2024, doi: 10.1007/s11042-024-19816-2.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," 2017, *arXiv*. doi: 10.48550/ARXIV.1706.09516.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[11] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.