INTERNATIONAL JOURNAL OF Telecommunications
AIR DEFENSE COLLEGE

An Ensemble Deep Learning-Based Approach for Microsatellite Instability Prediction in Gastrointestinal Cancer.

Youmna Ahmed radwan [1*], Hossam El- din Moustafa[2], Adel F. Moustafa[3], and Ahmed Saleh[4,]

[1] Computers Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura, Egypt; youmnaahmedradwan@mans.edu.eg
[2] Department of Electronics and Communications Engineering at the Faculty of Engineering, [3]Mansoura Uni-versity; hossam_moustafa@mans.edu.eg
Oncology Center, Mansoura University, Mansoura, Egypt, dr_adel_fathe@mans.edu.eg
Computers Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura, Egypt, aisaleh@mans.edu.eg
* Correspondence: youmnaahmedradwan@mans.edu.eg.

Citation: radwan, Y.; Moustafa, H.; Moustafa, A.; Saleh, A.
Inter. Jour. of Telecommunications, ***IJT'2025****, Vol. 05, Issue 02, pp. 1-16, 2025.*
Doi 10.21608/ijt.2025.412498.1131
Editor-in-Chief: Youssef Fayed.
Received: 09/08/2025.
Accepted date: 24/09/2025.
Published date: 24/09/2025.

Abstract: Microsatellite instability (MSI) is considered a significant biomarker for gastrointestinal (GI) cancer prognosis and treatment planning. Traditionally, molecular assays such as polymerase chain reaction (PCR) testing and immunohistochemistry (IHC) have been used to determine MSI status. Despite their effectiveness, these methods are labor-intensive and time-consuming .MSI tumors are known to respond better to immunotherapy due to their high mutational burden and increased immunogenicity, making accurate MSI assessment vital for selecting appropriate treatments. This study proposes a novel ensemble framework that combines Xception and InceptionResNetV2 using a soft-voting strategy to predict MSI directly from histopathological images. Unlike prior studies, which focused on single architectures or more complex ensembles, our approach integrates complementary CNN features with methodological simplicity and computational efficiency. The suggested ensemble model outperformed earlier methods with an accuracy of 96.97% and an area under the curve (AUC) of 99.57%. These results demonstrate the potential of efficient ensemble learning methods in advancing AI-assisted pathology, facilitating more personalized treatment decisions, and ultimately improving outcomes for patients receiving immunotherapy.

Keywords: Microsatellite instability, Gastrointestinal cancer, Transfer learning, ensemble model.

1. Introduction

Cancer continues to be a major global cause of mortality, with GI cancers making a substantial contribution to this global burden. All tumors that potentially impact the organs, such as the esophagus, stomach, liver, pancreas, colon, and rectum, are referred to as GI cancers[1]. According to the International Agency for Research on Cancer (IARC) under the World Health Organization (WHO), GI cancers account for about 1 in 4 cancer cases (26%) and 1 in 3 cancer-related deaths (35%) worldwide [2]. Among GI cancers, gastric and colorectal types occur frequently and have high fatality rates. Globally, colorectal cancer (CRC) ranks as the third most diagnosed cancer and the second most common cause of cancer-related deaths. Simultaneously, Gastric cancer (GC) also holds the position of being the fifth most diagnosed cancer and the fourth in terms of mortality rate. [3]. Alarmingly, the global impact of CRC is predicted to expand dramatically in the future years. By 2040, the number of new cancer cases is expected to increase dramatically by 63% to 3.2 million annually, while the

number of deaths may increase by 73% to 1.6 million [4].Consequently, there is an urgent need to increase initiatives that concentrate on prevention, early diagnosis, and enhanced treatment methods.

To combat GI cancer, early diagnosis is crucial. This can improve patient survival rates, overall quality of life, and treatment efficacy. Less intrusive therapeutic approaches will also be available. A multidisciplinary workup

including clinical examination, imaging, endoscopy, and histopathologic analysis of biopsy material is required to diagnose GI cancers [5]. Despite the existence of various diagnostic methods, several limitations persist, including the invasiveness and aggressive nature of some examinations, the time-consuming nature of many procedures, and an over-reliance on the individual physician's personal judgment and experience. Moreover, the biological complexity and heterogeneity of GI tumors pose challenges for early detection and personalized treatment. Recently, the role of MSI in pathophysiology has received more attention. With implications for prognosis and treatment response, particularly to immunotherapy, MSI stands out as a crucial biomarker [6].The evaluation of MSI has made molecular diagnostics indispensable in this context.

Microsatellites are short repetitive DNA sequences which can be error-prone during replication due to DNA polymerase slippage [7]. The Mismatch repair (MMR) system is responsible for correcting replication errors, thereby maintaining genomic stability. Deficiency in MMR (dMMR) occurs when critical MMR proteins, such as MLH1, MSH2, MSH6, or PMS2, are epigenetically silenced or have their genes mutated, causing microsatellite errors to accumulate—a phenomenon known as MSI.MSI results in a high mutational burden and contributes to the development of tumors with distinct molecular and immunogenic characteristics [8]. Microsatellite stable (MSS) tumors, on the other hand, possess a functional MMR system, which maintains genomic stability, exhibits a lower mutation burden. Microsatellite status is strongly associated with GI cancers, most notably CRC and GC, and plays a critical role in guiding treatment and improving patient outcomes. Table 1 summarizes the difference between the two types of tumors [9].

Table 1. Summary of differences between MSS and MSI tumors.

Feature	MSS tumors	MSI tumors
Mismatch Repair Status	Proficient Mismatch Repair	Deficient Mismatch Repair
Tumor Mutation Burden	Low	High (hypermutated)
Response to Immunotherapy	Poor response to immunotherapy	High response to immunotherapy
Associated Syndrome	Not typically associated with hereditary syndromes	Often associated with Lynch syndrome

The most common methods for determining MSI are IHC, which evaluates the presence or absence of MMR proteins, and PCR, which identifies changes in microsatellite length between tumor and normal tissues. More recently, MSI has also been assessed using next-generation sequencing (NGS) methods in conjunction with broader genomic profiling [10].

Identifying MSI tumors is clinically important because they often have a high mutational burden and unique histopathological features. Most importantly, MSI predicts improved responsiveness to immunotherapies, such as immune checkpoint inhibitors, which have altered the patients' treatment Options. Additionally, MSI testing is a crucial screening technique for Lynch syndrome, a genetic predisposition to GI cancers.

Traditional diagnostic techniques for the assessment of MSI face several challenges, including high costs, labor-intensive procedures, time-consuming processes, reliance on access to specialized molecular laboratories, and dependence on tissue quality. Emerging artificial intelligence (AI) technologies, particularly deep learning in digital pathology and medical imaging, provide solutions to these challenges. AI algorithms have the ability to integrate complex data from imaging, pathology, and clinical records to provide more comprehensive and personalized assessments. This can improve patient stratification, speed up early cancer detection, and facilitate the creation of more individualized and efficient treatment plans [11]. In GI oncology, AI algorithms can

quickly and effectively analyze Hematoxylin and Eosin (H&E) stained histopathology slides to predict MSI; their performance is frequently better than that of humans in recognizing patterns, as illustrated in Figure 1. This allows patients to receive appropriate treatment based on their cancer type, with MSI tumors responding better to immunotherapies, while MSS tumors respond better to chemotherapy [12].

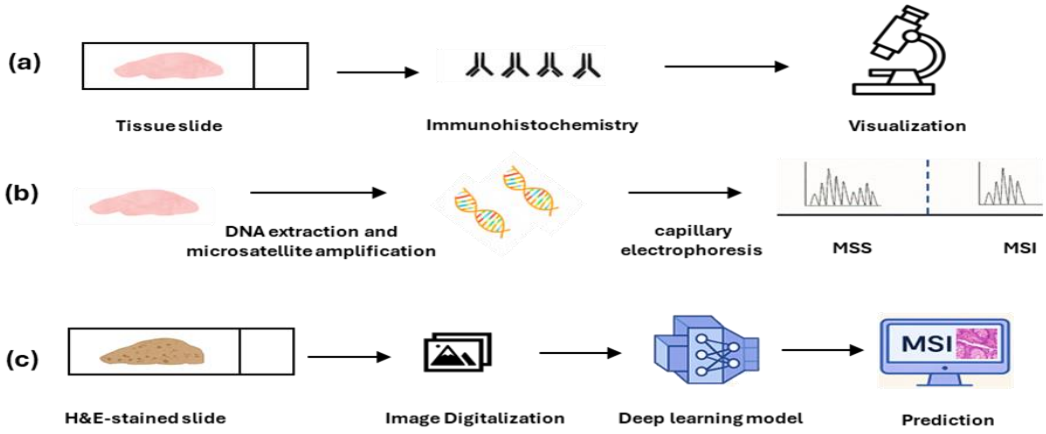

Figure 1. Detection of microsatellite instability (MSI) using traditional methods (a) IHC (b) PCR vs. (c) Modern method using AI techniques.

Although deep learning has shown significant promise in digital pathology, detecting microsatellite status from histopathological images remains a challenging task. A primary reason is the lack of clear morphological differences between MSI and MSS cases, which makes it inherently difficult to achieve high classification accuracy. Furthermore, dataset imbalance was frequently not sufficiently addressed in earlier research, which can result in biased models and a negative impact on model performance.

To address these challenges, this study makes the following contributions:

- Balanced learning through undersampling: To solve the MSI/MSS class imbalance, systematic undersampling was used to ensure equal representation of MSI and MSS cases and reduce bias in model training.
- Streamlined ensemble with efficient integration: A soft voting ensemble of two complementary CNNs (Xception and InceptionResNet-v2) avoids the complexity of stacked meta-learners by directly combining their prediction probabilities, leveraging different feature strengths. This probability-level fusion lowers the risk of overfitting, simplifies implementation, and reduces computational costs.
- Comparative benchmarking: Evaluated both models and their ensemble across accuracy, recall, precision, F1-score, and AUC for fair and consistent performance assessment.
- Improved performance: The suggested framework outperforms all prior works on the same dataset, exhibiting both robustness and clinical potential with an accuracy of 96.97% and an AUC of 99.57%.
- Clinical applicability: The approach emphasizes high accuracy alongside minimal false-positive and false-negative rates, ensuring reliable performance for real-world MSI screening.

2. Literature Review

Deep learning techniques have radically transformed the field of medical image analysis. In gastrointestinal (GI) cancer imaging, many studies have applied deep learning approaches across various imaging modalities such as CT, MRI, and endoscopy. However, a subset of research has focused on the Microsatellite status prediction from H&E histopathology images, which is a significant biomarker influencing treatment decisions in GI cancers. Recent deep learning models have demonstrated promising results in predicting MSI directly from histopathological images without the need for excessive lab tests. Most of the Prior studies have relied on a common and foundational dataset introduced by Kather et al.[13], which consists of histological image tiles derived from the Cancer Genome Atlas (TCGA) cohort. A subset of 192,312 histopathological images stained with H&E and preserved using formalin-fixed paraffin-embedding (FFPE) is publicly accessible via Kaggle and has become a benchmark resource for MSI prediction tasks. However, existing models still face limitations in

achieving high predictive accuracy and addressing the class imbalance between MSI and MSS cases. A summary of key studies utilizing this dataset is provided in Table 2.

Kather et al. [14] introduced a deep learning approach capable of predicting MSI directly from histopathology images in GI cancers, including CRC and GC. The authors applied a deep learning architecture (ResNet18) to identify MSI by analyzing histological patterns from TCGA without any molecular data. The German colorectal cancer cohort (DACHS) was used as an external validation. Achieving AUC values between 0.77 and 0.84 across several datasets, their model had excellent predictive performance.

Venkatesh et al. [15] proposed a modified ResNet architecture for the binary classification of MSI and MSS in GI cancer using histopathological images. The publicly accessible dataset of 192312 images on Kaggle was used for the study. The researchers evaluated baseline models such as logistic regression, a 4-layer feedforward neural network, and a CNN, followed by transfer learning using VGG16 and various versions of ResNet (ResNet-18, 34, 50, 101, 152). Their modified 41-layer ResNet model, which they developed based on these insights, performed the best of all the models, with an accuracy of 89.81% and an F1-score of 91.78%.

Khan and Loganathan [16] applied transfer learning techniques for the prediction of microsatellite status in GI cancer. They employed Xception, a convolutional neural network architecture renowned for its depthwise separable convolutions, which enhance parameter efficiency and reduce computational cost. Utilizing the dataset of 192312 images accessible on Kaggle, the model demonstrated its reliability with an AUC of 0.932 and a test accuracy of 90.17%.

Ghosh and Santosh [17] introduced a stacked generalization-based ensemble Deep Neural Network for binary classification of GI cancer histological images into MSI and MSS categories. Their framework enhanced classification performance by integrating predictions from three base models—a modified VGG16, DenseNet201, and a custom CNN—through a meta-learner to create an ensemble model. As a result, they were able to benefit from various feature extraction capabilities. Utilizing a dataset of 192,315 histological images accessible on Kaggle, they achieved an impressive accuracy and sensitivity of 94.91% and 95.95%, respectively, along with an AUC of 0.9821. This method demonstrates how ensemble models can enhance the diagnostic accuracy and dependability of automated pathology workflows.

Pamuk and Erikçi [18] proposed a deep learning approach to predict microsatellite instability in GI cancer from histopathology images. Their work utilized 150,000 image patches from a histological dataset available on Kaggle. They employed transfer learning to compare and evaluate nine pretrained CNN models. Among these, VGG19 achieved the highest classification results, recording an accuracy of 90.60%, a precision of 88.60%, and an AUC of 90.60%.

Table 2. Related studies in the classification of MSI and MSS tumors.

Study	year	Approach	Result	Advantages	Disadvantages
[14]	2019	CNN based on ResNet18	AUC: 0.77–0.84 (across several datasets)	• The paper provided a less invasive and scalable method for predicting MSI status from histological images. • Validated on external datasets.	• No other metrics were mentioned for evaluation • Low to Moderate performance • Imbalanced dataset
[17]	2021	Stacked ensemble (VGG16, DenseNet201, custom CNN).	Accuracy:94.91% Sensitivity:95.95% precision:93.35% AUC: 09821.	• Improved performance through ensemble learning, surpassing individual base models.	• Imbalanced dataset • High computational cost.

Table 2. Continued

Study	year	Approach	Result	Advantages	Disadvantages
[16]	2022	Transfer learning using Xception	Accuracy: 90.17% AUC:0.932.	• Transfer learning improved training efficiency. • Good AUC and test generalization	• Imbalanced dataset • Evaluation was limited to a narrow set of metrics.
[15]	2022	Modified ResNet (custom architecture with 41 layers)	Accuracy:89.81% F1-score:91.78%	• Improved feature learning through architectural modifications. • Implemented several CNN models and compared them with modified ResNet.	• Imbalanced dataset • The accuracy achieved is relatively low compared to other studies. • Reported only accuracy and F1-score.
[18]	2025	Comparative study of nine pre-trained models. (Top performer: VGG19).	VGG19 achieved: Accuracy: 90.60% Precision: 88.60% Recall: 93.10% AUC: 90.60%	• Evaluated multiple models on the same dataset. • Used a balanced dataset. • Reported comprehensive performance metrics	• The proposed model shows slightly lower accuracy compared to previous transfer learning studies.

3. Materials and methods

This section focuses on the materials and methodology used in the current study. A publicly available dataset of histological images of GI cancer was used. Several preprocessing techniques, such as resizing, normalization, and undersampling, were applied to these images to address the class imbalance and ensure input dimension consistency. An average ensemble model was constructed by combining two pretrained CNNs: Xception and InceptionResNetV2. These models were adapted to the task by removing their top layers and adding new custom ones. Finally, the performance of the proposed ensemble model was evaluated using a set of standard classification metrics. An outline of the proposed methodology is shown in Figure 2.

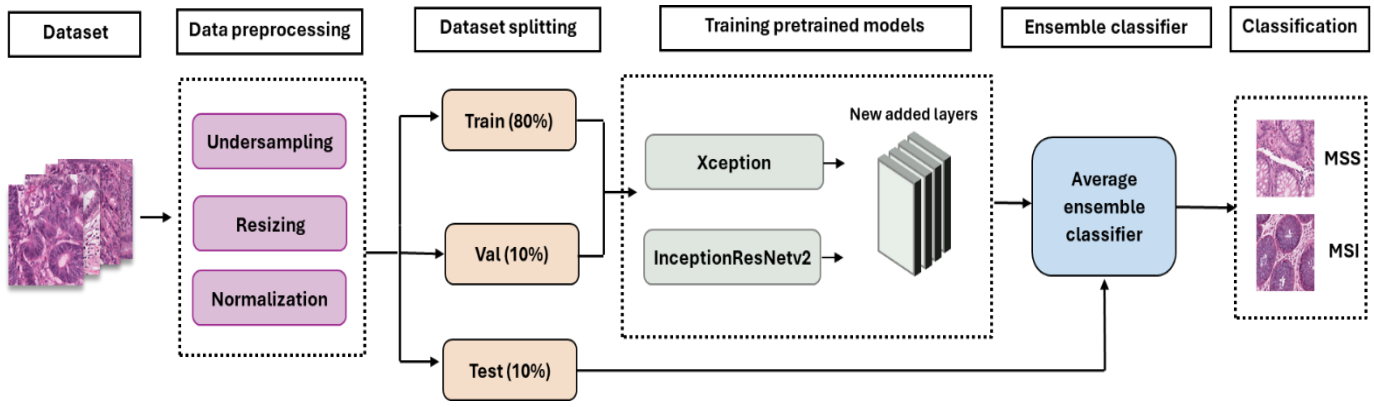

Figure 2. Key stages of the proposed methodology. The pipeline includes dataset preprocessing (undersampling, resizing, normalization), dataset splitting into training (80%), validation (10%), and testing (10%) sets, training of pretrained CNN models (Xception and InceptionResNetV2), integration via an average ensemble classifier, and final classification of histopathology images.

3.1. Dataset Description

In this study, histological images of CRC and GC were obtained from TCGA, which is a comprehensive and publicly accessible repository of cancer-related data [19]. The dataset consists of high-resolution whole-slide images (WSIs) stained with H&E, obtained from a population of patients, encompassing a wide variety of histopathological characteristics pertaining to CRC and GC. The original dataset included 411,890 distinct image patches that were taken from cancer patients' SVS-format WSIs from the TCGA cohort. During the dataset's creation, several preprocessing steps were carried out by Kather [13]. Every image was subjected to the same preprocessing pipeline, which included automatic tumor region detection, resizing to 224 × 224 pixels at a spatial resolution of 0.5 μm/px, and color normalization using the Macenko technique [20]. Based on the corresponding patient's microsatellite status, each image patch was annotated and classified as either MSS or MSI. A subset of 192,312 images—consisting of 117,273 MSS images and 75,039 MSIMUT images—was used in this study. This preprocessed and labeled dataset is publicly available on Kaggle [21] and Zendo [13]. Illustrative samples from the histological dataset are presented in Figure 3.

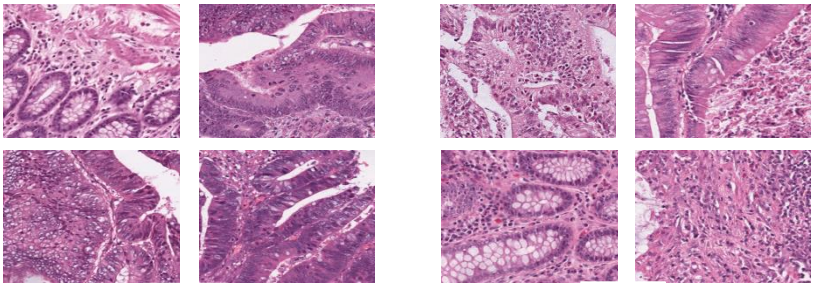

Figure 3. Samples of the histological dataset : (a) MSS, (b) MSI

3.2. Dataset Preparation and Preprocessing

The dataset went through a methodical preprocessing pipeline to ensure consistency and compatibility with deep learning models. The first step was to balance the dataset by undersampling the majority class, which produced an equal number of 75,000 images for the MSS and MSIMUT classes, as presented in Figure 4. This step was essential to prevent model bias towards the dominant class. Following class balancing, all images were resized to 299 × 299 pixels with three RGB channels to match the input dimensions required by the selected pretrained models (Xception and InceptionResNetV2). The next step was normalization, which involved scaling the pixel values ranging from 0 to 255 to a range of 0 to 1. This step was performed using Keras Image Data

Generator. The preprocessing techniques, undersampling, resizing, and normalization, collectively prepare the dataset in a form that is appropriate for deep learning classification tasks and computationally efficient. Finally, 80% of the dataset was used to train the proposed model, while 10% was set aside for validation, and another 10% for testing. Table 3 shows the image distribution across the train, validation, and test sets after applying undersampling.

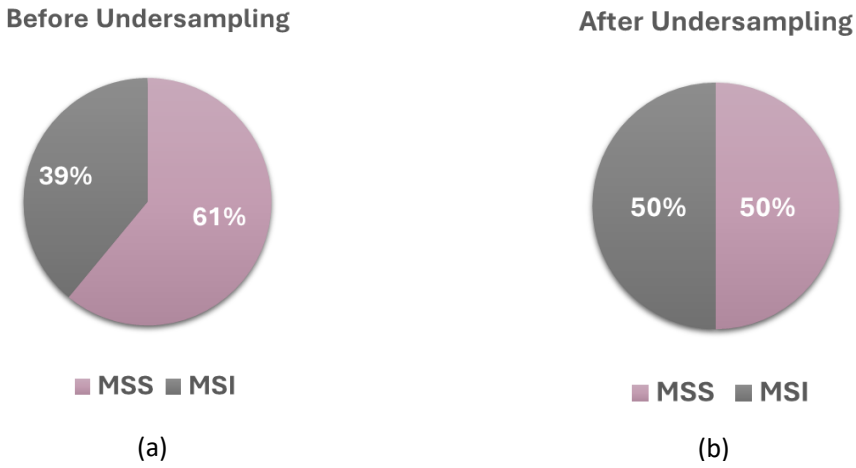

Figure 4. Distribution of dataset (a) before undersampling (b) after undersampling

Table 3. The number of images in each dataset split after undersampling.

Class Type	Train (80%)	Val (10%)	Test (10%)
MSS	60,000	15,000	15,000
MSI	60,000	15,000	15,000
Total	120,000	30,000	30,000

3.3. An overview of convolutional neural network

Convolutional Neural Network is a feedforward neural network that is designed to process visual data like images [22] .They are well-suited to perform tasks like object detection and image classification since they possess layers that automatically learn to identify features like edges, textures, and shapes.

Several deep layers with various tasks make up CNNs. These layers consist of the input layer, the convolution layer, the pooling layer, the fully connected layer (FC), and the final output layer. An image of size H × W × C is received by the input layer, where H and W represent the image's height and width, and C indicates the number of channels.

The convolution layer, which is the main feature extractor, applies several filters to identify significant patterns, such as edges and textures. A feature map highlighting some local features is generated by each filter [23]. To provide non-linearity and enable the model to identify complex patterns, a non-linear activation function—most frequently the Rectified Linear Unit (ReLU)—is added after each convolution. Mathematically, ReLU is expressed in equation (1).

$$f(x) = max(0, x) \#(1)$$

Where x is the output of the convolution operation.

Pooling layers are used to shrink the spatial dimensions of feature maps and reduce computational complexity. This process can involve methods like max pooling or average pooling, which downsample the input while retaining essential information. The feature maps that are produced after a series of convolution and pooling layers are passed into FC layers, which serve as classifiers by flattening the maps and linking every neuron to those in the preceding layer [24] .Finally, the CNN's classification or prediction is generated by the output layer. The choice of activation function in this layer can vary based on the specific task. In binary classification, a single neuron with a sigmoid activation function is often used, which generates a probability prediction between 0 and 1. The sigmoid function is defined by equation (2).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \#(2)$$

Where x denotes the function's input.

3.4. Pre-trained models and transfer learning approach

Transfer learning is a deep learning method where a model created for one task can be used as the starting point for another. Instead of building new deep neural networks from scratch, which require enormous volumes of labeled data and computational resources, transfer learning employs pre-trained models that have already been trained on huge datasets like ImageNet to identify useful features [25]. These pre-trained models, such as Xception and InceptionResNetV2, have learned rich feature representations that can generalize well to new visual tasks.

In this study, two pre-trained models, including Xception and InceptionResNetV2, were employed to classify histopathology images as MSI or MSS. To adapt them to our specific task, the original classification head was replaced with a custom set of new layers compatible with the desired output. These layers included a GlobalAveragePooling2D layer, Dropout layers for regularization with rates of 0.3 and 0.5, and a Dense layer using ReLU activation. The network concluded with a Dense layer employing a sigmoid activation function for binary classification. Figures 5 and 6 display the architecture of these models along with the newly added layers.

3.4.1 Inception-Resnet-v2

Inception-ResNet-v2, developed by Szegedy et al. [26] , is a CNN architecture that combines the strengths of two distinct deep learning models: Inception and ResNet. The model integrates Inception modules, which enable the network to use parallel convolutional layers with several kernel sizes (1×1, 3×3, and 5×5) to capture features at multiple scales, thus enhancing performance and efficiency. Meanwhile, it uses residual connections (also called skip connections) from the ResNet architecture, which bypass multiple layers, thereby preventing problems like vanishing gradients. Inception-ResNet-v2 merges these concepts by inserting residual connections into the Inception architecture, creating a hybrid model that is both deep and computationally efficient.
Inception-ResNet-v2 comprises approximately 55 million parameters and accepts inputs of size 299x299. The architecture achieves high accuracy on large-scale image classification tasks such as the ImageNet dataset. As illustrated in Figure 5 , the network is composed of three major components: the stem module, which applies several convolution and pooling layers to extract low-level features from the input image; Inception-ResNet-A/B/C blocks, which contain multiple parallel convolutional filters that extract features at various scales; and reduction blocks, which downsample feature maps while preserving representational power.

3.4.2 Xception

Xception, short for "Extreme Inception", is a deep CNN architecture introduced by François Chollet [27]. It is a modified version of Inception-V3. It is built upon the hypothesis that depthwise separable convolutions can replace the Inception modules, hence improving performance. The feature extraction base of the model is comprised of 36 convolutional layers organized in 14 modules. It processes 22.9 million parameters and accepts inputs of size 299x299.
In contrast to conventional convolution operations, Xception uses depth-wise separable convolutions to decouple spatial and cross-channel correlations. A depth-wise convolution (per-channel spatial filtering) is applied first, then a pointwise (1x1) convolution to combine the outputs across channels. This factorization improves learning efficiency and drastically lowers the number of parameters. As presented in Figure 6 , the structure of Xception has three main parts: the entry flow, which extracts basic features from the image; the

middle flow, which repeats the same block several times (8 times) to learn deeper patterns; and the exit flow, which finalizes the features before classification.

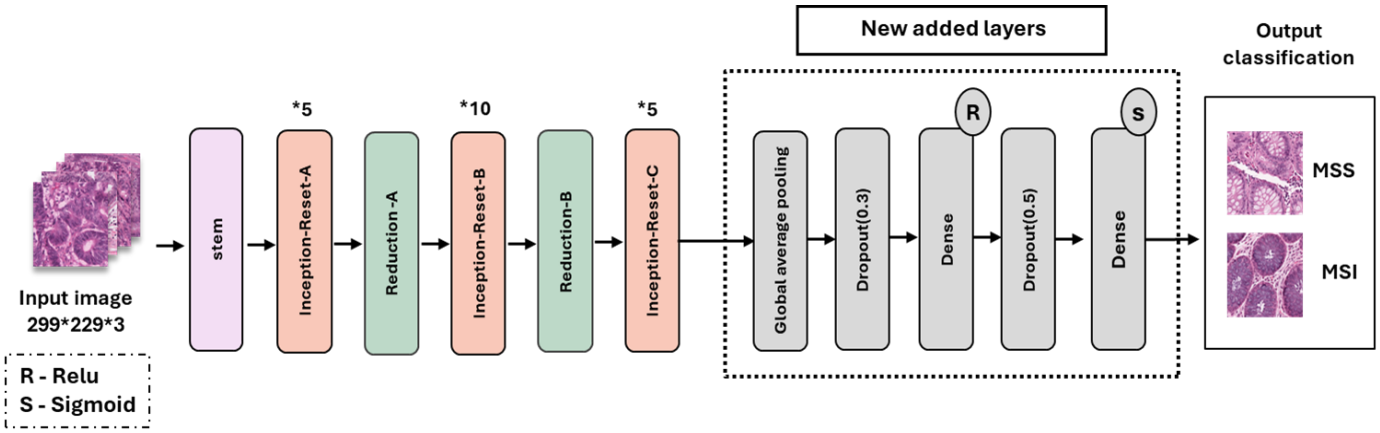

Figure 5. Inception-ResNet-v2 architecture with new added layers.

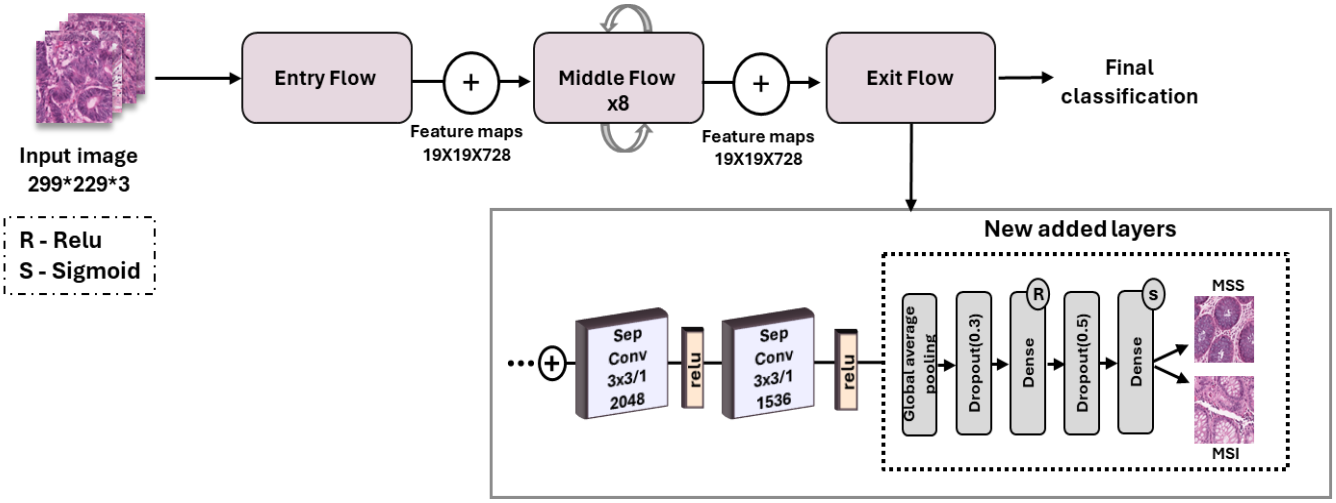

Figure 6. Xception architecture with new added layers.

The voting ensemble approach

Ensemble learning is a method that integrates several models, which are referred to as base learners or weak learners, to enhance overall performance and generalization. This approach produces more robust and accurate predictions by leveraging the strengths of different models. Bagging, boosting, and voting are the most popular types of ensemble learning. These methods vary in terms of model training and result combination.

Voting ensemble methods are widely used in classification problems to improve predictive performance by aggregating the outputs of multiple models [28].The main types of voting ensembles are hard voting and soft voting. In hard voting, each base classifier generates its prediction (a class label), and the final decision corresponds to the class receiving the majority vote. Soft voting, in contrast, averages the probability of outputs from all classifiers instead of relying on class labels. The final prediction is determined by identifying the class with the highest average probability. The average voting ensemble's predicted class is expressed by equation (3).

$$\hat{y}_i = argmax\left(\frac{1}{M}\sum_{j=1}^{M} p_{j(c)}\right) \#(3)$$

Where M refers to the number of classifiers and $p_{j(c)}$ represents the predicted probability for the class c as determined by the model j.

To enhance classification performance in determining microsatellite status from histopathological images, a soft voting ensemble classifier is proposed. In this approach, the final prediction is generated by averaging the predicted probabilities from each base classifier, which helps to leverage the strengths of multiple classifiers while reducing the variance associated with individual models. The prediction procedure based on the average of the two models' outputs is shown in the Figure 7.

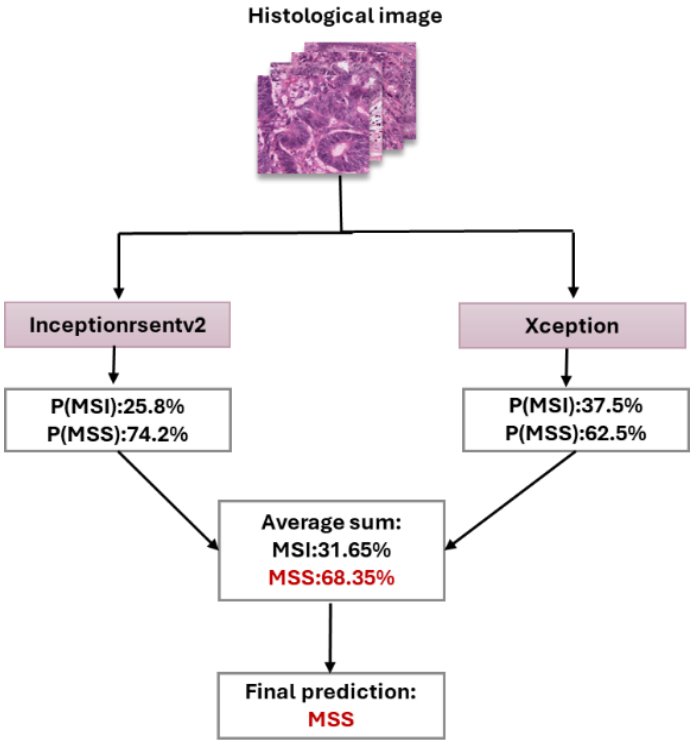

Figure 7. The average voting ensemble's class prediction process, where p refers to the prediction probabilities from each model expressed as percentages. The average of these probabilities is used to determine the final class.

3.5. Evaluation metrics

To evaluate the effectiveness of the proposed deep learning models, several performance measures have been employed in this study. A common key tool is the confusion matrix, which compares the predicted labels with the actual ground truth labels for each class to offer an overview of the prediction outcomes [29].It includes four key quantities:

- True Positives (TP): positive samples correctly classified as positive.
- True Negatives (TN): negative samples correctly classified as negative.
- False Positives (FP): negative samples incorrectly classified as positive.
- False Negatives (FN): positive samples incorrectly classified as negative

Based on these quantities, multiple metrics—such as accuracy, sensitivity, precision, and F1-score—can be computed. Accuracy (Acc) reflects the proportion of correctly predicted samples (both positive and negative) among all predictions. Precision (Pre) expresses the percentage of predicted positives that are actually correct. Recall (Rec) is the proportion of actual positives that the model successfully identifies. The F1-score computes

the harmonic mean of precision and recall, providing a balanced measure. The mathematical formulas for these metrics are given in Equations (4)–(7).

$$Accuracy = \frac{TN + TP}{FP + FN + TP + TN} \#(4)$$

$$Precision = \frac{TP}{FP + TP} \#(5)$$

$$Recall = \frac{TP}{FN + TP} \#(6)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \#(7)$$

Furthermore, the Receiver Operating Characteristic (ROC) curve, which compares the true positive rate against the false positive rate at various threshold values, is used to assess the classification model's effectiveness. An overall indicator of the classifier's performance is provided by the Area Under the Curve (AUC), where values approaching 1.0 suggest that it is powerful for distinguishing between classes.

4. Experimental results

This study was conducted in two phases. In the first phase, two pretrained CNN architectures were individually trained and evaluated. In the second phase, a soft voting ensemble classifier was employed to integrate the strengths of these models and improve overall predictive accuracy.

4.1. Performance evaluation of pre-trained CNN models

Two deep CNN architectures, InceptionResNet-v2 and Xception, were employed as base models. Training was performed using stochastic gradient descent (SGD), with a 0.01 learning rate and a batch size of 64. With an accuracy of 96.11% and an AUC of 99.32%, the Xception model outperformed the InceptionResNet-v2 model. It also achieved the highest recall of 96.17% and F1-score of 96.12%, indicating strong generalization. However, its precision (96.06%) was slightly lower than that of InceptionResNet-v2, which achieved a higher precision of 96.16%, along with a competitive accuracy of 95.66% and AUC of 99.31%. These results demonstrated that both models are well-suited for the task, with Xception being the most effective as a standalone classifier. However, further improvements were necessary to minimize false predictions, leading to the exploration of ensemble approaches.

4.2. Performance evaluation of the voting ensemble Approach

To enhance generalization and improve classification performance, the predictions of the individual models were integrated using a soft voting ensemble approach. This ensemble, proposed as the final model, performed the best overall, with an accuracy of 96.97%, a recall of 96.84%, a precision of 97.10%, an F1-score of 96.97%, and an AUC of 99.57%. These results indicate a well-balanced and robust classifier that effectively leverages the strengths of its base models. Table 4 summarizes the performance metrics of all evaluated models.

Confusion matrices and ROC curves were generated for all models and are shown in Figures 8 and 9, respectively. The confusion matrix for the proposed average ensemble demonstrated minimal false positives and false negatives, confirming its reliability in distinguishing between class labels.

All models' ROC curves showed AUC values above 99%, demonstrating strong discriminative power and reliable generalization ability. Notably, the suggested average ensemble had the highest AUC value of 99.57%, highlighting the effectiveness of ensemble strategies in improving classification performance.

Table 4. Evaluation metrics of all conducted experiments.

Model	Accuracy	Precision	Recall	F1-score	AUC
Inceptionresnet-v2	95.66%	96.16%	95.12%	95.64%	99.31%
Xception	96.11%	96.06%	96.17%	96.12%	99.32%
Soft voting (Average)	**96.97%**	**97.10%**	**96.84%**	**96.97%**	**99.57%**

5. Discussion

The study involved a two-phase approach for MSI/MSS status prediction from histopathological images. In the first Phase, two pre-trained CNN models (InceptionResNet-v2 and Xception) were employed by replacing the final classification layers to suit the binary classification problem. Both models showed strong performance on the dataset, with Xception achieving the highest results among them, reaching an accuracy of 96.11% and an AUC of 99.32%. However, residual misclassifications in the confusion matrix suggested that more improvement was required.

In the second phase, a soft voting ensemble technique was applied to improve robustness and reduce predictive variance. This approach significantly enhanced performance by averaging the predicted probabilities, effectively leveraging the complementary strengths of the base models. The resulting ensemble demonstrated superior performance, achieving an accuracy of 96.97%, a precision of 97.10%, a recall of 96.84%, an F1-score of 96.97%, and an exceptional AUC of 99.57%, consistently outperforming individual models across all evaluation metrics. Furthermore, the low number of misclassifications in the confusion matrix (only misclassifying 237 out of 7500 MSI cases and 217 out of 7500 MSS cases), as presented in Figure 8, demonstrates the ensemble model's applicability for real-world medical applications where both false positives and false negatives have serious consequences.

Notably, the Xception model alone achieved very strong results, which can be attributed to its use of depthwise separable convolutions. Histopathological images benefit greatly from the effective and fine-grained feature extraction made possible by this design. Nevertheless, the proposed ensemble consistently provided marginal but reliable improvements by combining complementary representations from both Xception and Inception-ResNetV2. Even modest gains in accuracy and AUC are clinically meaningful, since they translate into fewer misclassified patients and thus more reliable treatment planning. In diagnostic applications, a 1–2% improvement can correspond to dozens of correctly identified cases in large screening programs, which is highly impactful in practice.

Beyond predictive performance, interpretability is essential for clinical applicability. Deep learning models, including our ensemble, are often described as 'black boxes which could prevent their adoption in pathology workflows. Grad-CAM, SHAP, and LIME are examples of Explainable AI (XAI) techniques that could be used to identify regions of interest in histopathological images and uncover the characteristics that influence MSI/MSS predictions. Transparency would be increased, and pathologists could use their own knowledge to cross-validate AI-generated insights. Importantly, since MSI/MSS classification is often visually indistinguishable on H&E slides, such explanations could reveal patterns not detectable by the human eye. The adoption of AI in routine diagnostic procedures could be accelerated by reducing the gap between algorithmic performance and clinical trust through the integration of predictive performance and interpretability.

A detailed comparison of recent studies on MSI prediction using deep learning methods is summarized in Table 2, highlighting the performance and limitations of prior approaches. Although Kather et al. [14] demonstrated a CNN based on ResNet18 and reported an AUC range of 0.77–0.84 across multiple datasets; they didn't evaluate the model using important measures like F1-score, precision, and recall. Ghosh and Santosh [17] introduced a stacked generalization ensemble with multiple networks, reaching an accuracy of 94.91%, and an AUC of 0.9521. Even though they showed better performance through their ensemble model, the meta-learner architecture meant that they needed a lot of processing power and didn't specifically address the dataset imbalance. Both Khan et al. [16] and Venkatesh et al. [15] reported accuracy of 93.18% and 89.81%, respectively, by utilizing transfer learning techniques (using Xception and modified ResNet). However, neither study reported essential performance metrics such as precision and recall.

Pamuk and Ertürk [18] performed a comparative evaluation of pretrained CNN models, where VGG19 yielded a recall of 93.10%, a precision of 85.80%, a classification accuracy of 90.60%, and an AUC value of 90.60%. Although they used a balanced dataset and reported more detailed metrics, their overall performance was lower than recent ensemble or transfer learning methods.

In contrast, the proposed average ensemble in this study offers both state-of-the-art performance and practical applicability, outperforming all prior works. It demonstrates an approximate 2% improvement in accuracy over the best prior approach and delivers consistently high precision and sensitivity, supported by notably low false-positive and false-negative rates. These underscore the model's effectiveness and reliability as a robust solution for MSI/MSS classification in real-world medical applications.

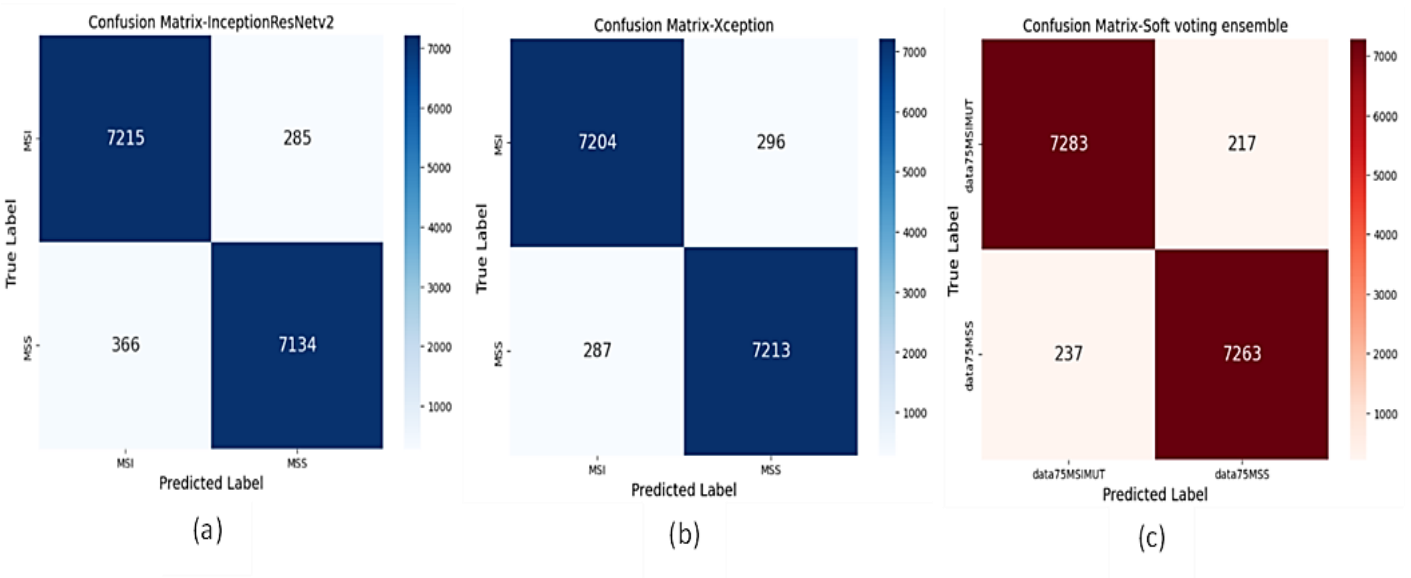

Figure 8. The confusion matrix of evaluated models. (a) Inceptionresnetv2, (b) Xception, (c) Soft voting.

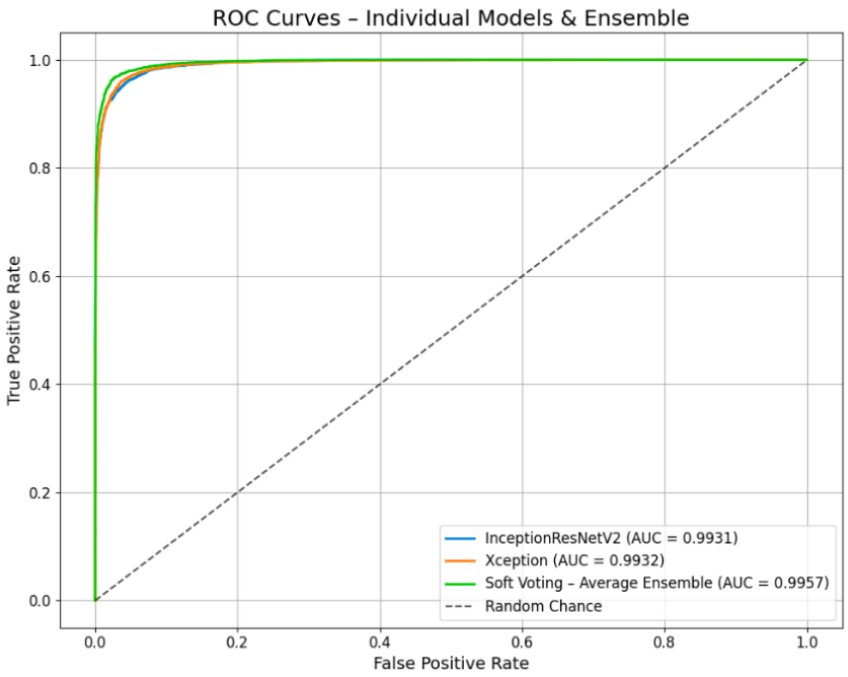

Figure 9.ROC curves of all experiments.

6. Conclusion and future work

Early and precise detection of MSS and MSI in GI cancer patients is essential for determining the optimal treatment plan and improving overall clinical outcomes. MSI is considered an important biomarker that helps identify patients who are likely to benefit from immunotherapy. Precise identification of MSI allows clinicians to personalize treatment plans, reduce unnecessary side effects, enhance therapeutic response rates, and ultimately improve patient survival and quality of life.

In this study, an average ensemble of pre-trained CNNs (InceptionResNet-v2 and Xception) is proposed to automatically identify MSI and MSS from histopathological images. With an accuracy of 96.97% and an AUC of 99.57%, the ensemble classifier successfully utilized the complementary strengths of the individual networks. These results demonstrated low false-positive and false-negative rates, confirming the reliability of the proposed model while surpassing the performance of individual learners and existing approaches. The strong overall performance of the classifier highlights its potential as a useful clinical decision-support tool to help oncologists and pathologists make prompt and accurate treatment decisions for patients with GI cancer.

Future research could explore different CNN architectures, experiment with alternative ensemble approaches like stacking and boosting, and validate the proposed model on external datasets to ensure its generalizability in different clinical settings. Importantly, prospective validation and cross-cohort evaluation across multiple institutions will be necessary to confirm robustness and ensure applicability in real-world clinical scenarios. Furthermore, integrating explainable AI techniques—such as SHAP or LIME—could improve interpretability and transparency, boost clinical confidence, and facilitate integration into actual diagnostic settings.

7. References

[1] S. Kuntz et al., "Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review," European Journal of Cancer, vol. 155, pp. 200–215, Sep. 2021, doi: 10.1016/J.EJCA.2021.07.012.

[2] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] M. Arnold et al., "Global Burden of 5 Major Types of Gastrointestinal Cancer," Gastroenterology, vol. 159, no. 1, pp. 335-349.e15, Jul. 2020, doi: 10.1053/j.gastro.2020.02.068.

[4] E. Morgan et al., "Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN," Gut, vol. 72, no. 2, pp. 338–344, Feb. 2023, doi: 10.1136/gutjnl-2022-327736.

[5] A. S. Almasoud, M. Maray, H. K. Alkahtani, F. A. Alotaibi, M. M. Alnfiai, and A. Sayed, "Gastrointestinal Cancer Detection and Classification Using African Vulture Optimization Algorithm with Transfer Learning," IEEE Access, vol. 12, pp. 23122–23131, 2024, doi: 10.1109/ACCESS.2024.3351773.

[6] J. Wang et al., "Mutational analysis of microsatellite-stable gastrointestinal cancer with high tumour mutational burden: a retrospective cohort study," The Lancet Oncology, vol. 24, no. 2, pp. 151–161, Feb. 2023, doi: 10.1016/S1470-2045(22)00783-5.

[7] M. Ratti, A. Lampis, J. C. Hahne, R. Passalacqua, and N. Valeri, "Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches," Cellular and Molecular Life Sciences, vol. 75, no. 22, pp. 4151–4162, Nov. 2018, doi: 10.1007/s00018-018-2906-9.

[8] R. Gupta, S. Sinha, and R. N. Paul, "The impact of microsatellite stability status in colorectal cancer," Current Problems in Cancer, vol. 42, no. 6, pp. 548–559, Nov. 2018, doi: 10.1016/J.CURRPROBLCANCER.2018.06.010.

[9] K. Heinimann, "Toward a Molecular Classification of Colorectal Cancer: The Role of Microsatellite Instability Status," Frontiers in Oncology, vol. 3, 2013, doi: 10.3389/fonc.2013.00272.

[10] F. Zito Marino et al., "Microsatellite Status Detection in Gastrointestinal Cancers: PCR/NGS Is Mandatory in Negative/Patchy MMR Immunohistochemistry," Cancers, vol. 14, no. 9, p. 2204, Apr. 2022, doi: 10.3390/c cers14092204.

[11] S. Shafi and A. v. Parwani, "Artificial intelligence in diagnostic pathology," Diagnostic Pathology, vol. 18, no. 1, p. 109, Oct. 2023, doi: 10.1186/s13000-023-01375-z.

[12] M. Kreidieh, D. Mukherji, S. Temraz, and A. Shamseddine, "Expanding the Scope of Immunotherapy in Colorectal Cancer: Current Clinical Approaches and Future Directions," BioMed Research International, vol. 2020, no. 1, Jan. 2020, doi: 10.1155/2020/9037217.

[13] J. N. Kather, "Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples". Zenodo, Feb. 07, 2019. doi: 10.5281/zenodo.2530835.

[14] J. N. Kather et al., "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," Nature Medicine, vol. 25, no. 7, pp. 1054–1056, Jul. 2019, doi: 10.1038/s41591-019-0462-y.

[15] C. H. Sai Venkatesh, C. Meriga, M. G. V. L. Geethika, T. Lakshmi Gayatri, and V. B. K. L. Aruna, "Modified ResNetModel for MSI and MSS Classification of Gastrointestinal Cancer," Lecture Notes in Electrical Engineering, vol. 853, pp. 273–282, 2022, doi: 10.1007/978-981-16-9885-9_23.

[16] Z. Khan and R. Loganathan, "Transfer learning based classification of MSI and MSS gastrointestinal cancer," International journal of health sciences, vol. 6, no. S1, pp. 1857–1872, Mar. 2022, doi: 10.53730/ijhs.v6nS1.4952.

[17] S. Ghosh and K. C. Santosh, "Improved Gastrointestinal Screening: Deep Features using Stacked Generalization," in 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, Jun. 2021, pp. 196–201. doi: 10.1109/CBMS52027.2021.00071.

[18] Z. Pamuk and H. Erikçi, "A Comparative Analysis of Deep Learning Models for Prediction of Microsatellite Instability in Colorectal Cancer," Sakarya University Journal of Computer and Information Sciences, vol. 8, no. 1, pp. 136–151, Mar. 2025, doi: 10.35377/saucis...1638424.

[19] A. Janowczyk, "Download TCGA Digital Pathology Images (FFPE) - Andrew Janowczyk." Accessed: Jun. 28, 2025. [Online]. Available: https://andrewjanowczyk.com/download-tcga-digital-pathology-images-ffpe.

[20] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE, Jun. 2009, pp. 1107–1110. doi: 10.1109/ISBI.2009.5193250.

[21] "TCGA COAD MSI vs MSS Prediction (JPG)." Accessed: Jun. 28, 2025. [Online]. Available: https://www.kaggle.com/datasets/joangibert/tcga_coad_msi_mss_jpg

[22] V. A. Hiremani and K. K. Senapati, "Quantifying apt of RNN and CNN in Image Classification," Lecture Notes in Electrical Engineering, vol. 748, pp. 721–733, 2021, doi: 10.1007/978-981-16-0275-7_59.

[23] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Progress in Artificial Intelligence, vol. 9, no. 2, pp. 85–112, Jun. 2020, doi: 10.1007/s13748-019-00203-0.

[24] Y. Dogan, "A New Global Pooling Method for Deep Neural Networks: Global Average of Top-K Max-Pooling," Traitement du Signal, vol. 40, no. 2, pp. 577–587, Apr. 2023, doi: 10.18280/ts.400216.

[25] M. Iman, H. R. Arabnia, and K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," Technologies, vol. 11, no. 2, p. 40, Mar. 2023, doi: 10.3390/technologies11020040.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, pp. 4278–4284, Feb. 2017, doi: 10.1609/AAAI.V31I1.11231.

[27] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[28] K. Akyol, E. Uçar, Ü. Atila, and M. Uçar, "An ensemble approach for classification of tympanic membrane conditions using soft voting classifier," Multimedia Tools and Applications, vol. 83, no. 32, pp. 77809–77830, Feb. 2024, doi: 10.1007/s11042-024-18631-z.

[29] E. Sherbiny et al., "A Diabetes Mellitus Prediction Model Based on Supervised Machine Learning Techniques," International Journal of Telecommunications, vol. 05, no. 01, pp. 1–11, Mar. 2025, doi: 10.21608/IJT.2025.359269.1083.