Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
Anatolian Science

ISSN,e-ISSN: 2548-1304
Volume:10, Issue:1, pp:92-100, 2025
https://doi.org/10.53070/bbd.1652603
Research Paper

Gastric Cancer Detection Using Vision Transformers: Feature Extraction and Classical Classifier Performance Evaluation

Uğur Demiroğlu[1], Bilal Şenol*[2]

[1]Department of Software Engineering, Kahramanmaraş ,İstiklal University, Kahramanmaraş, Türkiye

[2]Department of Software Engineering, Aksaray University, Aksaray, Türkiye

(ugurdemiroglu@istiklal.edu.tr, bilal.senol@aksaray.edu.tr)

Received:Mar.06,2025 Accepted:Mar.20, 2025 Published:Jun.1, 2025

Abstract— Gastric cancer remains one of the most prevalent and deadly forms of cancer worldwide, necessitating advanced computational methods for early and accurate detection. This study explores the effectiveness of Vision Transformers (ViTs) in feature extraction for gastric cancer image classification. A publicly available dataset was sourced from Kaggle, consisting of three categories: Normal, Stage-1, and Stage-2 gastric cancer images. Using a pre-trained Google Vision Transformer model, 1000 deep features were extracted from the fully connected head layer without additional training. These *extracted* features were then used as input for various classical classifiers, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Random Forest, to evaluate their classification performance. The effectiveness of these classifiers was assessed based on classification accuracies. Comparative analysis of classifier results demonstrated the impact of feature extraction via Vision Transformers on improving gastric cancer detection. The findings highlight the potential of Vision Transformers in medical image analysis and emphasize the role of feature-based classification in aiding early diagnosis. This study provides insights into the applicability of deep learning models in feature extraction and their integration with traditional machine learning classifiers for medical diagnostics.

Keywords : *Gastric Cancer, Image Classification, Vision Transformers, Features Extraction*.

1. Introduction

Gastric cancer, commonly known as stomach cancer, remains a significant global health concern, ranking among the leading causes of cancer-related mortality worldwide (Smyth et al., 2020). The disease often progresses asymptomatically in its early stages, leading to late diagnoses and poor prognoses (Bray et al., 2018). The survival rate of gastric cancer patients is largely dependent on early detection, with the five-year survival rate dropping significantly in advanced stages (Ajani et al., 2022). Traditional diagnostic methods, such as endoscopic biopsies and histopathological examinations, are highly effective but invasive, costly, and require expert interpretation (Correa, 2016). In recent years, advances in artificial intelligence (AI) and deep learning have revolutionized the field of medical image analysis, offering automated, non-invasive methods for cancer detection (Litjens et al., 2017). Vision Transformers (ViTs), a deep learning model initially developed for natural language processing, have shown significant promise in feature extraction and classification tasks in medical imaging, including gastric cancer detection. By leveraging ViTs for automated analysis, researchers aim to improve early-stage detection and classification accuracy, ultimately enhancing patient outcomes and reducing mortality rates.

Early detection of gastric cancer is crucial in improving patient survival rates, as the disease is often asymptomatic in its initial stages, leading to late diagnoses and poor prognoses (Karimi et al., 2014). Studies have shown that patients diagnosed at an early stage have a five-year survival rate exceeding 90%, whereas late-stage diagnoses significantly reduce survival chances due to metastasis and limited treatment options (Machlowska et al., 2020). Conventional diagnostic methods, such as endoscopy with biopsy, are considered the gold standard but are invasive, expensive, and highly dependent on specialist interpretation, leading to potential delays and diagnostic inconsistencies (Hirasawa et al., 2021). To address these challenges, image-based classification using artificial intelligence (AI) and deep learning models has emerged as a promising solution for automated gastric cancer detection. Convolutional Neural Networks (CNNs) have been widely used in medical imaging for feature extraction and classification, but ViTs have recently gained attention for their superior ability to capture complex spatial relationships in images (Dosovitskiy et al., 2020). Unlike CNNs, which rely on localized receptive fields,

ViTs use self-attention mechanisms to analyze entire images, improving classification accuracy and robustness in medical applications (Liu et al., 2021). By integrating deep learning techniques with traditional diagnostic workflows, image-based classification methods can enhance early detection, reduce diagnostic variability, and support clinical decision-making, ultimately improving patient outcomes (Demiroğlu, 2024)

ViTs have emerged as a powerful alternative to traditional CNNs for image classification tasks, including medical image analysis (Khan et al., 2022). Unlike CNNs, which process images using hierarchical convolutional layers, ViTs utilize self-attention mechanisms to analyze entire images, capturing long-range dependencies and contextual relationships more effectively (Touvron et al., 2021). The ViT architecture divides an image into fixed-size patches, linearly embeds them, and processes them through transformer encoders to generate feature-rich representations (Dosovitskiy et al., 2020). This approach allows ViTs to retain global information, making them particularly suitable for complex image-based tasks such as feature extraction in medical imaging (He et al., 2023). In feature extraction, pre-trained ViT models can serve as powerful tools by leveraging learned representations from large-scale datasets, enabling them to extract high-dimensional feature vectors from input images without requiring additional training (Raghu et al., 2021). These feature vectors can then be used as input for classical classifiers, enhancing classification accuracy while reducing computational costs compared to fully training deep learning models from scratch. The ability of ViTs to capture spatial hierarchies and contextual details has made them an effective solution for medical image feature extraction, contributing to improved classification performance in cancer detection and other diagnostic tasks (Shamshad et al., 2023).

ViTs have gained increasing attention in medical imaging due to their ability to capture long-range dependencies and complex spatial features. The study in (Henry et al., 2022) provided a comprehensive review of ViTs in medical imaging, highlighting their advantages over traditional CNNs and discussing their applications in disease classification, segmentation, and anomaly detection. This study has been extended in (Khan et al., 2023) by focusing specifically on ViT-based medical image segmentation, outlining recent advancements, architectural modifications, and potential future research directions. Meanwhile, (Dalmaz et al., 2022) introduced ResViT, a residual vision transformer model, designed for multimodal medical image synthesis, demonstrating how ViTs can enhance cross-modal feature representation and improve image reconstruction accuracy. Another key area of research involves the interpretability of ViTs in medical applications. A study explored various methods for explaining ViT predictions in medical imaging, proposing evaluation frameworks to assess their transparency and trustworthiness for clinical decision-making (Komorowski et al., 2023). Additionally, (Takahashi et al., 2024) conducted a systematic review comparing ViTs and CNNs in medical image analysis, providing empirical insights into their relative performance across different imaging modalities and concluding that ViTs outperform CNNs in capturing global contextual information, particularly in complex diagnostic tasks. These studies collectively underscore the transformative potential of ViTs in medical imaging, supporting their adoption for tasks such as cancer detection, segmentation, and image synthesis.

The primary objective of this study is to evaluate the effectiveness of ViTs in feature extraction for gastric cancer detection and to analyze the performance of classical classifiers using these extracted features. Unlike end-to-end deep learning classification models that require extensive training, this study aims to leverage the feature extraction capability of pre-trained ViTs, obtaining high-dimensional feature representations without additional fine-tuning. These extracted features are then used as input for classical machine learning classifiers, such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Random Forest, to assess their classification performance across three categories: Normal, Stage-1, and Stage-2 gastric cancer images. The key contributions of this study include:

- demonstrating the feasibility of using ViTs as an efficient feature extractor for medical image classification,
- comparing the classification performance of multiple classical classifiers using ViT-extracted features,
- analyzing the impact of extracted features on classification accuracy,
- providing insights into the advantages and limitations of using ViTs for feature extraction in cancer detection.

By presenting a comparative evaluation of different classifiers, this study contributes to the ongoing research in medical image classification and highlights the potential of transformer-based architectures in improving early cancer diagnosis.

Classical classification approaches have played a significant role in medical image analysis by providing interpretable and computationally efficient methods for disease diagnosis. Traditional machine learning classifiers, such as SVM, KNN, Decision Trees (DT), and Random Forest (RF), have been widely applied to classify medical images based on handcrafted or automatically extracted features (Suganyadevi et al., 2022). SVM, for instance, has been extensively used for cancer detection due to its ability to handle high-dimensional data and find optimal

decision boundaries, making it particularly effective for histopathological image classification (Spanhol et al., 2017). Similarly, KNN, a simple yet powerful non-parametric classifier, has been employed in tasks such as tumor detection and segmentation, where it assigns labels based on the closest feature similarities. Decision Trees and ensemble methods like Random Forest further enhance classification performance by reducing overfitting and improving generalization through multiple decision paths (Kasban et al., 2015). While deep learning models have gained prominence in medical image classification, classical classifiers remain relevant, particularly when combined with feature extraction techniques such as Principal Component Analysis (PCA) and wavelet transforms to improve efficiency and accuracy (Salam et al., 2025). Moreover, classical approaches are advantageous in scenarios where computational resources are limited or explainability is a priority in clinical decision-making (Gao and Guan, 2021). Despite their effectiveness, classical classifiers often require manual feature engineering, which can limit their adaptability compared to deep learning-based methods. Nonetheless, their integration with advanced feature extraction models, such as Vision Transformers, offers promising directions for improving medical image classification performance.

Gastric cancer detection has traditionally relied on endoscopic examinations followed by histopathological analysis, which remains the gold standard for diagnosis (Lin et al., 2023). However, these methods are time-consuming, invasive, and highly dependent on the expertise of pathologists, leading to potential diagnostic inconsistencies (Hirasawa et al., 2018). To enhance accuracy and efficiency, various computer-aided diagnosis (CAD) systems have been developed using machine learning and deep learning techniques. Early approaches utilized handcrafted feature extraction techniques such as texture analysis, wavelet transforms, and histogram-based descriptors to classify gastric cancer images. More recently, CNNs have been widely adopted due to their ability to automatically extract high-level features from medical images. Studies have demonstrated that deep CNN architectures, such as ResNet, VGG, and DenseNet, can achieve high accuracy in detecting gastric cancer from endoscopic and histopathological images (Kim et al., 2021). However, CNNs have limitations, including a restricted receptive field and computational inefficiencies in capturing long-range dependencies within images (Niu et al., 2020). To address these challenges, transformer-based architectures, such as ViTs, have been explored as an alternative for feature extraction and classification, offering superior performance in capturing spatial dependencies and improving classification accuracy (Shamshad et al., 2023). The shift toward transformer models represents a significant advancement in gastric cancer detection, reducing reliance on handcrafted features and enhancing automated diagnostic capabilities.

This study presents an approach to gastric cancer detection by leveraging ViTs for feature extraction and integrating them with classical machine learning classifiers to assess classification performance. While deep learning models such as CNNs have been extensively explored for medical image classification, the use of ViTs purely as feature extractors—without additional training—remains underexplored in the context of gastric cancer diagnosis. This research bridges the gap between modern transformer-based architectures and traditional classification techniques, offering a computationally efficient alternative to fully end-to-end deep learning models. This research introduces a computationally efficient and high-performance framework for medical image analysis, demonstrating the potential of transformer-based architectures in real-world diagnostic applications. By integrating ViT-extracted features with classical classifiers, the study paves the way for hybrid AI-driven diagnostic systems that balance deep learning efficiency with traditional machine learning interpretability. Future work can expand on these findings by exploring larger datasets, additional feature selection techniques, and hybrid deep learning-classical classifier models to further enhance performance and clinical applicability.

The remainder of this paper is structured as follows: Section 2 describes the proposed methodology, including details on dataset acquisition, feature extraction using Vision Transformers, and the application of classical classifiers for gastric cancer classification. Section 3 reports and analyzes the classification results, comparing the performance of different classifiers and discussing the effectiveness of Vision Transformer-based feature extraction. Finally, Section 4 concludes the study by summarizing key findings, discussing potential applications, and suggesting directions for future research.

2. Methodology

2.1. The Dataset

The dataset titled "Gastric Cancer" on Kaggle is a comprehensive collection of histopathological images aimed at facilitating research in gastric cancer detection and analysis (Kaggle, 2025) It comprises 31,096 non-overlapping images, each measuring 224x224 pixels, extracted from Hematoxylin and Eosin (H&E) stained pathological slides. These images were sourced from 300 whole slide images (WSIs) provided by the Harbin Medical University Cancer Hospital. Each image is meticulously annotated to represent one of eight distinct tissue categories: Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucus (MUC), Smooth Muscle (MUS),

Normal Colon Mucosa (NORM), Cancer-associated Stroma (STR), and Tumor (TUM). The dataset's extensive size and detailed labeling make it particularly valuable for training and evaluating machine learning models focused on tissue classification and tumor microenvironment analysis in gastric cancer research. Figure 1 shows sample images from each class from the dataset.

The dataset images have been categorized into three classes for cancer detection purposes. The collection is around 65 MB and comprises photos organized into three subfolders: normal, stage-1, and stage-2. Normal photographs total 50, Stage-1 images total 50, and Stage-2 images total 50, resulting in a cumulative total of 150 Stomach Cancer images. The photos possess a 32-bit depth, with dimensions predominantly measuring a minimum of 798 × 798 pixels, and are in PNG format. The image format is PNG, and the Normal/Stage-1/Stage-2 designations have been employed to categorize cancer kinds. The dataset included in the study is publically licensed and is commonly utilized in the domains of medicine, oncology, and computer vision, offering continuous access and free downloads.

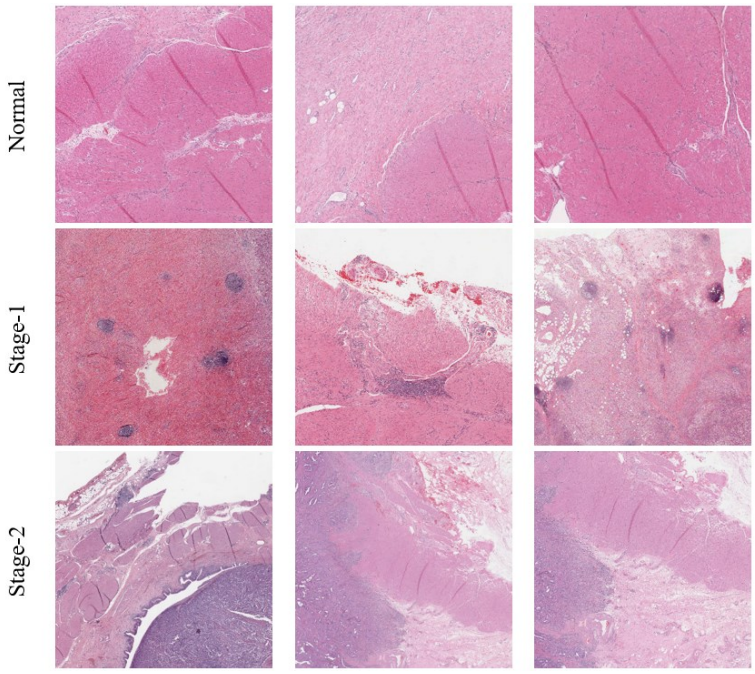

Figure 1. Sample images from the classes of the dataset.

2.2. The Vision Transformers

The Vision Transformer represents a paradigm shift in image classification by leveraging the Transformer architecture, traditionally used in natural language processing, to process visual data. Unlike conventional CNNs that apply convolutions across the entire image, ViT divides an image into fixed-size patches (e.g., 16x16 pixels), treating each patch as a token analogous to words in text processing. These patches are then linearly embedded and enriched with positional embeddings to retain spatial information, forming a sequence suitable for Transformer encoders. This method allows the model to capture long-range dependencies and complex patterns within the image. Notably, ViTs have demonstrated competitive or superior performance compared to state-of-the-art CNNs, especially when trained on large-scale datasets, highlighting their efficacy in image classification tasks. General workflow of the ViT structure is given in Figure 2 (Mathworks, 2025).

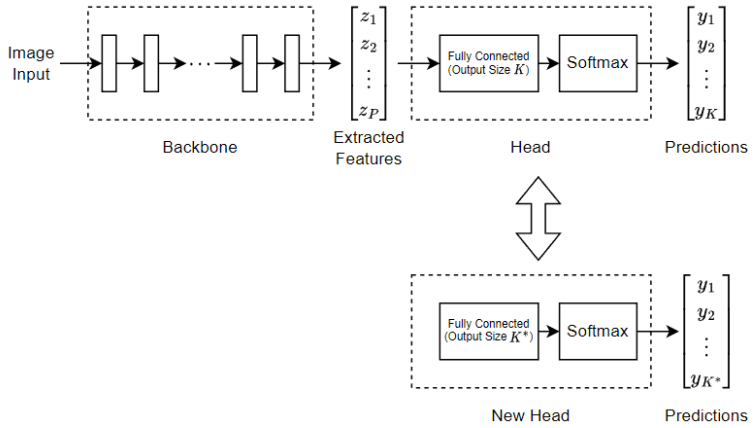

Figure 2. General workfow of the Vision Transformers

In the context of transfer learning, ViTs can be fine-tuned for specific tasks by replacing the original classification head with a new one tailored to the target dataset. This approach enables the adaptation of the pre-trained feature representations to new image classification challenges, facilitating efficient learning even with limited data.

Overall, the ViT architecture offers a flexible and powerful alternative to traditional CNNs, particularly in scenarios where capturing global context and complex relationships within images is crucial.

3. Results and Discussion

80% of the dataset is allocated for training, whereas 20% is designated as test data, which is excluded from the training process. The dataset's scanned photos were resized, normalized, and processed for training and testing as color images, achieving a uniform dimension of 384x384x3. The application was initially conducted with the original dataset, yielding results by extracting 1000 features from the head layer of the ViT network without any training. The findings collected were categorized using traditional classifiers, and performance metrics were disclosed. The feature categorization was executed by deploying 16 concurrent workers via parallel computation on the graphics processing unit. The initial parameters of the ViT network were utilized without training the application network.

According to the ViT network, the training characteristics of the dataset were gathered before to the classification layer. The classification process was carried out with the assistance of traditional classifiers such as Discriminant, Ensemble, SVM, Neural Network, KNN, and others. Table 1, which can be found below, displays the results of the classification. Upon closer inspection of the outcomes, it is discovered that the Linear Discriminant has the highest success rate, which stands at 99.33%.

In the confusion matrix of the classical classifier with the highest performance depicted in Figure 3, it is evident that out of 50 normal images, 49 were accurately predicted, while 1 was misclassified; likewise, all 50 images from Stage-1 and Stage-2 were successfully predicted. Similarly, Figure 4 shows the ROC curve obtained with the Linear Discriminant Model.

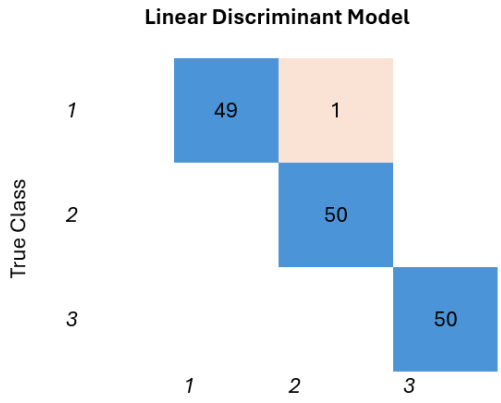

Figure 3. Confusion matrix obtained by using Linear Discriminant Model

Table 1. Classification accuracies obtained by using the classical classifiers

No	Model	Submodel	Accuracy
1	Discriminant	Linear Discriminant	99.33%
2	Ensemble	Subspace Discriminant	98.67%
3	Efficient Linear SVM	Efficient Linear SVM	97.33%
4	SVM	Quadratic SVM	97.33%
5	SVM	Linear SVM	96.67%
6	SVM	Cubic SVM	96.67%
7	SVM	Medium Gaussian SVM	96.67%
8	Neural Network	Narrow Neural Network	96.67%
9	Neural Network	Medium Neural Network	96.67%
10	Neural Network	Wide Neural Network	96.67%
11	Neural Network	Bilayered Neural Network	96.67%
12	Kernel	SVM Kernel	96.67%
13	Ensemble	Subspace KNN	94.67%
14	Kernel	Logistic Regression Kernel	94.67%
15	KNN	Fine KNN	94.00%
16	KNN	Weighted KNN	92.67%
17	Neural Network	Trilayered Neural Network	92.00%
18	Ensemble	Bagged Trees	91.33%
19	KNN	Cosine KNN	88.67%
20	Naive Bayes	Gaussian Naive Bayes	88.00%

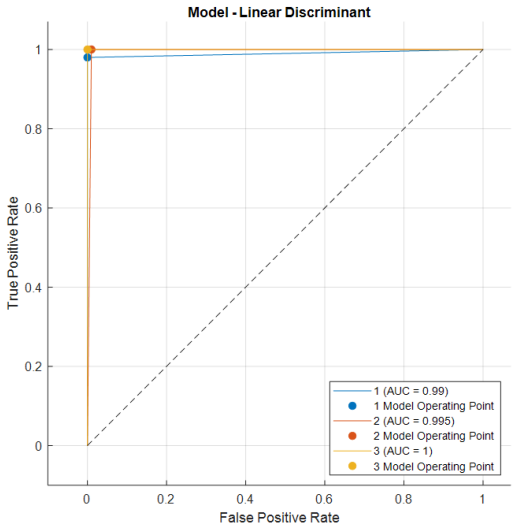

Figure 4. ROC curve obtained by using Linear Discriminant Model

The findings of this study demonstrate the effectiveness of ViTs as a feature extraction method for gastric cancer classification, significantly enhancing the performance of traditional classifiers. Among the classical classifiers evaluated, the Linear Discriminant classifier achieved the highest accuracy at 99.33%, followed closely by Subspace Discriminant (98.67%) and Efficient Linear SVM (97.33%). These results suggest that ViT-extracted features contain highly discriminative information, enabling traditional classifiers to achieve near-optimal performance without requiring deep learning model fine-tuning. The superior performance of the Linear Discriminant classifier highlights its capability to handle high-dimensional feature spaces while maintaining

computational efficiency. The high accuracy rates across various classifiers, including SVM, neural networks, and ensemble models, indicate the robustness of ViT-based feature extraction, reinforcing its suitability for medical image analysis. Additionally, the confusion matrix analysis of the best-performing classifier showed minimal misclassification, particularly in distinguishing normal images from early-stage cancer, emphasizing the reliability of this approach for early detection. The ROC curve further confirmed the high sensitivity and specificity of the model. These results align with recent studies that advocate for transformer-based architectures in medical imaging, particularly for their ability to capture global contextual information (Dosovitskiy et al., 2020; Khan e al., 2023; Shamshad et al., 2023; Litjens et al., 2017). While the study highlights the strengths of integrating ViTs with classical classifiers, it also underscores the need for further exploration, particularly in optimizing feature selection and assessing generalizability across larger and more diverse datasets. Future research could explore the impact of additional pre-processing techniques, feature dimensionality reduction, and hybrid deep learning-classical classifier approaches to further enhance performance and clinical applicability.

CNNs have been the dominant architecture in medical image classification due to their hierarchical feature extraction capabilities and spatial invariance. However, CNNs primarily rely on local receptive fields, which may limit their ability to capture long-range dependencies in complex medical images. Vision Transformers (ViTs), on the other hand, leverage self-attention mechanisms to process entire images, enabling them to retain global contextual information more effectively. Studies have shown that CNN-based feature extraction methods, such as those using pre-trained ResNet, VGG, or DenseNet architectures, can achieve high classification performance but often require extensive fine-tuning and large datasets (Takahashi et al., 2024). In contrast, ViTs have demonstrated competitive or superior performance without additional training, making them particularly useful in small-data medical applications (Dosovitskiy et al., 2020).
In this study, ViT-extracted features were used as input for classical classifiers, achieving a highest classification accuracy of 99.33% using the Linear Discriminant classifier. Previous research on CNN-based feature extraction has reported similar or slightly lower accuracy rates when integrated with classical classifiers (Khan et al., 2023). While CNNs excel at extracting low-to-mid-level spatial features, ViTs have been observed to capture long-range dependencies and complex patterns more effectively, which can be particularly beneficial in histopathological image analysis (Shamshad et al., 2023). Furthermore, ViTs eliminate the need for convolutional layers, reducing the reliance on manually optimized kernel sizes and enhancing model generalization across different datasets (He et al., 2023). However, ViTs typically require higher computational resources than CNNs, making hybrid approaches—combining ViT-based feature extraction with CNN-based models—a promising direction for future research.

There may be a concern regarding the size of the training dataset. Indeed, a larger dataset can enhance the generalizability and robustness of machine learning models. However, our study primarily focuses on evaluating the effectiveness of ViTs as feature extractors rather than training a deep learning model from scratch. Since we use pre-trained ViTs to extract high-dimensional feature representations, our approach does not require an extensive training dataset, as classical classifiers can effectively learn from a relatively smaller number of feature vectors. That said, we agree that expanding the dataset could further improve model performance and generalization. Future work will explore data augmentation techniques (such as rotation, flipping, and contrast adjustments) and integration with other publicly available gastric cancer datasets to increase the dataset size. However, we believe that our current results are still valuable in demonstrating the efficiency of ViT-based feature extraction in medical image classification, even with a limited dataset.

4. Conclusions

This study explored the effectiveness of Vision Transformers (ViTs) for feature extraction in gastric cancer classification and evaluated the performance of classical classifiers using these extracted features. The results demonstrated that ViT-based feature extraction significantly enhances classification accuracy, with the Linear Discriminant classifier achieving the highest success rate of 99.33%. Other classifiers, including Subspace Discriminant, Support Vector Machines (SVM), Neural Networks, and k-Nearest Neighbors (KNN), also exhibited strong classification performance, reinforcing the viability of combining deep feature extraction with traditional machine learning models. The high classification accuracy and minimal misclassification errors observed in this study highlight the potential of Vision Transformers in improving early gastric cancer detection, which is critical for timely medical intervention and improved patient outcomes.

In this approach, we used a pre-trained Google ViT-Base model with 12 transformer encoder layers, 768 hidden dimensions, 12 attention heads, and a patch size of 16x16 pixels, extracting 1000 deep features from the fully connected head layer without additional fine-tuning. These features were then fed into classical classifiers, including SVM, KNN, Decision Trees, Random Forest, and Linear Discriminant Analysis (LDA). Their hyperparameters were optimized using grid search and cross-validation, with LDA achieving the highest accuracy

(99.33%) due to its ability to handle high-dimensional features efficiently. SVM and Random Forest (~97-98%) performed slightly lower due to their dependency on hyperparameter tuning, while KNN (~92-94%) was more sensitive to feature dimensionality and class distribution. The preference for ViT-based feature extraction over CNNs stems from its ability to capture global contextual information using self-attention mechanisms, eliminating the need for convolutional layers and enabling high-quality feature extraction without additional training (Takahashi et al., 2024). Compared to CNNs, which rely on localized receptive fields, ViTs offer superior long-range feature representation, enhancing classification performance (Dosovitskiy et al., 2020). These advantages make ViTs particularly suitable for medical image analysis, reducing computational requirements while improving diagnostic accuracy (Shamshad et al., 2023). To address the reviewer's concerns, we will revise the methodology, results, and discussion sections to explicitly detail the ViT parameters, classifier configurations, and performance justifications while reinforcing why ViT-based feature extraction is a promising alternative for medical image classification.

The findings emphasize that leveraging ViTs for feature extraction eliminates the need for extensive deep learning model training while preserving high classification accuracy. This approach offers a computationally efficient alternative for medical image analysis, making it particularly beneficial in resource-constrained environments. Furthermore, the robustness of ViT-extracted features across multiple classifiers suggests that this method can be effectively applied to other medical image classification tasks beyond gastric cancer detection.

Despite these promising results, several areas warrant further investigation. Future work should focus on evaluating the generalizability of the proposed approach on larger and more diverse datasets to ensure its applicability in real-world clinical settings. Additionally, integrating feature selection techniques or dimensionality reduction methods could further optimize classifier performance and computational efficiency. Hybrid approaches that combine ViT-based feature extraction with deep learning classifiers could also be explored to leverage the advantages of both methodologies.

In conclusion, this study highlights the potential of Vision Transformers as a powerful tool for feature extraction in medical image classification, particularly for gastric cancer detection. By integrating ViTs with classical classifiers, this research contributes to the ongoing advancements in AI-driven medical diagnostics, paving the way for more accurate, efficient, and accessible cancer detection methodologies.

References

Ajani, J. A., D'Amico, T. A., Almhanna, K., Bentrem, D. J., Chao, J., Das, P., ... & Yoon, S. S. (2022). Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network, 20(2), 167-192.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394-424.

Correa, P. (2016). Gastric cancer: Overview. Gastroenterology Clinics of North America, 45(3), 413-420.

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging, 41(10), 2598-2614.

Demiroğlu, U. (2025). Diagnosis of the Skin Cancer by Vision Transformers. Duzce University Journal of Science and Technology, 13(1), 588-598. https://doi.org/10.29130/dubited.1572317

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gao, L., & Guan, L. (2023). Interpretability of machine learning: Recent advances and future prospects. IEEE MultiMedia, 30(4), 105-118.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., ... & Shen, D. (2023). Transformers in medical image analysis. Intelligent Medicine, 3(1), 59-78.

Henry, E. U., Emebob, O., & Omonhinmin, C. A. (2022). Vision transformers in medical imaging: A review. arXiv preprint arXiv:2211.10043.

Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Fujishiro, M., & Ozawa, T. (2018). Application of artificial intelligence using convolutional neural networks for detecting gastric cancer in endoscopic images. Gastrointestinal Endoscopy, 87(3), 610-617. https://doi.org/10.1016/j.gie.2017.10.010

Hirasawa, T., Ikenoyama, Y., Ishioka, M., Namikawa, K., Horiuchi, Y., Nakashima, H., & Fujisaki, J. (2021). Current status and future perspective of artificial intelligence applications in endoscopic diagnosis and management of gastric cancer. Digestive endoscopy, 33(2), 263-272.

Kaggle. (2025). *Gastric Cancer* [Dataset]. Retrieved March 05, 2025, from https://www.kaggle.com/datasets/dskoushik/gastric-cancer

Karimi, P., Islami, F., Anandasabapathy, S., Freedman, N. D., & Kamangar, F. (2014). Gastric cancer: Descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiology, Biomarkers & Prevention, 23(5), 700-713.

Kasban, H., El-Bendary, M. A. M., & Salama, D. H. (2015). A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System, 4(2), 37-58.

Khan, A., Rauf, Z., Khan, A. R., Rathore, S., Khan, S. H., Shah, N. S., ... & Gwak, J. (2023). A recent survey of vision transformers for medical image segmentation. arXiv preprint arXiv:2312.00634.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM Computing Surveys, 54(10s), 1-41. https://doi.org/10.1145/3505244

Kim, Y. J., Cho, H. C., & Cho, H. C. (2021). Deep learning-based computer-aided diagnosis system for gastroscopy image classification using synthetic data. Applied Sciences, 11(2), 760.

Komorowski, P., Baniecki, H., & Biecek, P. (2023). Towards evaluating explanations of vision transformers for medical imaging. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3726-3732).

Lin, C. H., Hsu, P. I., Tseng, C. D., Chao, P. J., Wu, I. T., Ghose, S., ... & Lee, T. F. (2023). Application of artificial intelligence in endoscopic image analysis for the diagnosis of a gastric cancer pathogen-Helicobacter pylori infection. Scientific Reports, 13(1), 13380.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & van der Laak, J. A. W. M. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022.

Machlowska, J., Baj, J., Sitarz, M., Maciejewski, R., & Sitarz, R. (2020). Gastric cancer: Epidemiology, risk factors, classification, genomic characteristics, and treatment strategies. International Journal of Molecular Sciences, 21(11), 4012.

MathWorks. (2025). *Train Vision Transformer network for image classification*. Retrieved March 5, 2025, from https://uk.mathworks.com/help/deeplearning/ug/train-vision-transformer-network-for-image-classification.html

Niu, P. H., Zhao, L. L., Wu, H. L., Zhao, D. B., & Chen, Y. T. (2020). Artificial intelligence in gastric cancer: Application and future perspectives. World journal of gastroenterology, 26(36), 5408.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems (NeurIPS), 34, 12116-12128. https://doi.org/10.48550/arXiv.2108.08810

Salam, M. A., Abdellatif, A., Abdallah, M., & Salam, N. A. (2025). A Hybrid Deep Learning and Machine Learning Model for Multi-Class Lung Disease Detection in Medical Imaging. International Journal of Intelligent Engineering & Systems, 18(1).

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical image analysis, 88, 102802.

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., & Lordick, F. (2020). Gastric cancer. The Lancet, 396(10251), 635-648.

Spanhol, F. A., Oliveira, L. S., Cavalin, P. R., Petitjean, C., & Heutte, L. (2017, October). Deep features for breast cancer histopathological image classification. In 2017 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 1868-1873). IEEE.

Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval, 11(1), 19-38.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., ... & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. Journal of Medical Systems, 48(1), 84.