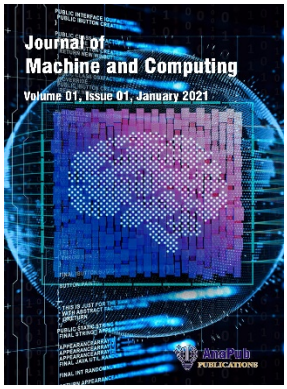
Journal Pre-proof



Machine Learning-Driven Feature Extraction and Dimensionality Reduction for Gastric Cancer Image Classification

Angati Kalyan Kumar, Gangadhara Rao Kancharla

DOI: 10.53759/7669/jmc202404052

Reference: JMC202404052

Journal: Journal of Machine and Computing.

Received 23 October 2023; Revised form 02 March 2024; Accepted 22 May 2024.

Please cite this article as:

Angati Kalyan Kumar, Gangadhara Rao Kancharla, “Machine Learning-Driven Feature Extraction and Dimensionality Reduction for Gastric Cancer Image Classification”, Journal of Machine and Computing. (2024). Doi: https:// doi.org/10.53759/7669/ jmc202404052

This PDF file contains an article that has undergone certain improvements after acceptance. These enhancements include the addition of a cover page, metadata, and formatting changes aimed at enhancing readability. However, it is important to note that this version is not considered the final authoritative version of the article.

Prior to its official publication, this version will undergo further stages of refinement, such as copyediting, typesetting, and comprehensive review. These processes are implemented to ensure the article's final form is of the highest quality. The purpose of sharing this version is to offer early visibility of the article's content to readers.

Please be aware that throughout the production process, it is possible that errors or discrepancies may be identified, which could impact the content. Additionally, all legal disclaimers applicable to the journal remain in effect.

© 2024 Published by AnaPub Publications.

Machine Learning-Driven Feature Extraction and Dimensionality Reduction for Gastric Cancer Image Classification

[1] Angati Kalyan Kumar, [2] Gangadhara Rao Kancharla
[1,2] Department of Computer Science & Engineering, University College of Sciences, Acharya Nagarjuna University, Guntur,Andhra Pradesh, India
[1] kalyank442@gmail.com, [2] kancherla123@gmail.com

ArticleInfo
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmcXXXXXXXX
Received xx December xxxx; Revised form xx December xxxx; Accepted xx December xxxx
Available online xx January xxxx.
©2022 The Authors. Published by AnaPub Publications.
This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Abstract - Cancer is the leading cause of death globally, affecting various organs in the human body. Early diagnosis of gastric cancer is essential for improving survival rates. However, traditional diagnosis methods are time-consuming, require multiple tests, and rely on specialist availability. This motivates the development of automated techniques for diagnosing gastric cancer using image analysis. While existing computerized techniques have been proposed, challenges remain. These include difficulty distinguishing healthy from cancerous regions in images and extracting irrelevant features during analysis. This research addresses these challenges by proposing a novel deep learning-based method for gastric cancer classification. The method utilizes deep feature extraction, dimensionality reduction, and classification techniques applied to a gastric cancer image dataset. This approach achieves high accuracy (99.32%), sensitivity (99.13%), and specificity (99.64%) in classifying gastric cancer.

Keywords - Gastric cancer, Feature extraction, Inception, Classification, Support vector machine.

I. INTRODUCTION

Gastric cancer, a widespread cancer that develops in the stomach's epithelial cells, holds the unfortunate distinction of being the fourth most common cancer globally and the second leading cause of cancer-related deaths[1]. More than 1 million gastric cancer diagnoses resulted in over 768,000 deaths worldwide in 2020, solidifying its position as the fifth most lethal cancer globally[2]. Even with surgery, chemotherapy, and radiation, the chances of surviving advanced stomach cancer for five years remain below 30%[3]. Over 95% of gastric cancers are adenocarcinomas, classified by their location in the stomach and the type of cells involved. Interestingly, the incidence of this specific cancer type shows significant geographic variation, with higher rates observed in Asia, Africa, South America, and Eastern Europe[4]. While endoscopy and surgery are currently the main tools for diagnosing stomach cancer, some patients, particularly those in rural settings, may avoid them due to concerns about discomfort and affordability [5]. In essence, feature selection streamlines the disease prediction process for AI. It ensures the AI has the most relevant information, leading to faster, more accurate diagnoses and a deeper understanding of the disease itself [6]. Unveiling diseases from mountains of patient data can overwhelm AI. Autoencoders are the answer – acting like data ninjas, they shrink this information, highlighting key features like a detective. This magic lies in a special neural network that compresses data into a "latent space," like a zip file containing only crucial details for disease prediction. By focusing on these essentials, autoencoders empower AI for faster, more accurate diagnoses, paving the way for a deeper understanding of diseases [7][8]. Machine learning is revolutionizing disease diagnosis. Researchers are developing new methods to analyze data and identify the most important factors for predicting, detecting, and predicting the course of diseases [9] [10]. The field of medical imaging is experiencing a surge in the use of artificial intelligence (AI), with research suggesting AI models can significantly improve diagnostic accuracy.

Authors Pre-Proof

These models achieve results on par with, or even exceeding, those of clinical experts in certain circumstances[11]. Modified logistic regression model [12] and OLGV3 Net Classifier [13] can be used in Cancer disease classification. This advancement allows for the application of AI algorithms to multi-dimensional data, including clinical and follow-up information, conventional images (endoscopy, histopathology, and CT scans), and molecular biomarkers. This approach has the potential to improve risk surveillance for gastric cancer in patients with established risk factors[14]. Machine learning allows computers to learn from data and experience, enabling them to automatically adapt and improve their performance on a given task. Deep learning (DL) is a special kind of machine learning that uses layers like building blocks to uncover more and more meaningful details from raw data, step by step[15]. A CNN model was developed to enhance the ability to efficiently differentiate between early gastric cancer and noncancerous lesions, achieving outstanding diagnostic accuracy[16][17]. The two networks leveraged transfer learning. A pre-trained VGG-16 network, trained on the vast ImageNet image classification dataset, provided the foundation for these models. This approach effectively initialized and optimized the weights within the new networks[18]. The system relies on an SVM algorithm to categorize the images as either normal or containing cancer[19]. The following research gaps are identified in the existing work and well addressed in this research paper. **1.Limited Generalization** refers to a model's inability to perform well on new, unseen data because it was trained on a dataset that lacks sufficient diversity and variability. This means the model may work effectively on the training data but fails to accurately predict outcomes on different datasets, reducing its practical applicability in real-world scenarios. **2.Mitigating Overfitting** involves applying strategies to ensure that a machine learning model does not perform exceptionally well only on training data but also generalizes well to new, unseen data. **3.** The "**Complexity and Computational Efficiency Problem**" arises from the effectiveness of autoencoders in dimensionality reduction, as they introduce significant computational complexity and overhead. This research gap indicates the need for more efficient dimensionality reduction techniques that can maintain essential information without compromising the computational efficiency, enabling faster and more scalable processing of gastric cancer images.

II. LITERATURE REVIEW

This section provides a detailed overview of existing techniques for gastric cancer classification Afrash et al.[20] used six machine learning models, including SVM(RBF), XGBoost, SVM Linear, Random Forest, Multilayer Perceptron, and KNN, to predict the early risk of gastric cancer based on lifestyle factors. Among these models, the XGBoost model performed best, achieving an accuracy of 83.41%. Cheng-Mao Zhou et al.[21] applied six machine learning algorithms to predict total gastric cancer deaths after surgery: Logistic Regression (LR), Gradient Boosting Machine (GBM), Gradient Boosted Decision Trees (GBDT), Random Forest, Tree-based Regression (Tr), and Extreme Gradient Boosting Classifier (XGBC). The highest accuracy achieved by the LR algorithm is 75.9%. However, this study has several limitations. Firstly, it is a retrospective study of postoperative gastric cancer patients. Secondly, the model's drawback is its inability to predict outcomes for other malignant tumors. Finally, the prediction variations observed may not be applicable to gastric cancer patients with other malignant tumors. Talebi et al.[22] employed machine learning algorithms to predict metastasis in gastric cancer patients. Support Vector Machine (SVM) achieved an accuracy of 93%. Functional and enrichment analyses were conducted using Gene Ontology (GO) and the Kyoto Database of Genes. However, a limitation of this study is that the validation is not up to par. Yoon HJ et al.[23] introduced a Visual Geometry Group (VGG)-16 model for the classification of endoscopic images as early gastric cancer (EGC) and for predicting depth. The overall accuracy achieved was 73%. However, a limitation of the study is that it does not analyze the accuracy of EGC detection specifically. Huang B et al.[24] introduced the GastroMIL algorithm for diagnosing gastric cancer, which attained an accuracy of 92%. However, a limitation of this study is that the survival time of gastric cancer individuals from the TGCA cohort varied.

Sakai Y et al.[25] proposed a deep learning system using convolutional neural networks to automatically detect early gastric cancer in endoscopic images. This system leveraged transfer learning with a limited dataset containing two classes (cancerous and normal) and focused on detailed texture information of lesions. While achieving an accuracy of 87.6%, the approach faced limitations. Firstly, the system over-detected certain regions with very irregular surface textures, potentially leading to false positives. Secondly, it struggled with out-of-focus or deeper lesions, resulting in missed detections. Zheng X et al.[26] introduced a deep convolutional neural network, named EBVNet, designed to predict Epstein-Barr virus-associated gastric cancer (EBVaGC) from histopathology images. EBVNet achieved an impressive average Area Under the Receiver Operating Curve (AUROC) of 96.9% during internal cross-validation. Notably, EBVaGC exhibits a strong response to immune checkpoint inhibitor treatments.

Existing research identifies critical gaps in gastric cancer diagnosis, including validation, over-detection, and misclassification issues. Limited generalization to diverse patient populations, coupled with a lack of detailed feature analysis, hinders model refinement. Dataset constraints impact generalization, while mitigating overfitting is crucial for model robustness. Addressing complexity in dimensionality reduction techniques is essential for scalable processing.

Authors Pre-Proof

III. PROPOSED METHODOLOGY

III.i) Traditional Inception Model

The traditional Inception model, also known as GoogLeNet, was introduced in the paper "Going Deeper with Convolutions" by Szegedy et al. It comprises several key architectural elements designed to achieve both depth and computational efficiency. The model begins with an input layer where images of varying sizes are fed into the network. Successive layers consist of convolutional operations with different filter sizes (e.g., 1x1, 3x3, 5x5) and strides, followed by Rectified Linear Unit (ReLU) activation functions. Inception modules, a central component of the architecture, incorporate parallel convolutional operations with various filter sizes, concatenating feature maps along the depth dimension. Max pooling layers interspersed between convolutional layers downsample feature maps and reduce spatial dimensions. Fully connected layers follow, culminating in a final layer with a softmax activation function for producing class probabilities. This traditional Inception model pioneered the development of deep convolutional neural networks for image classification tasks, emphasizing depth and computational efficiency through its innovative architectural design.

III.ii) Proposed Methodology

a) Dataset

The first dataset is the gastric cancer dataset. It consists of 245,196 images. These images can be classified into two categories: normal and abnormal as shown in figure 1. There are 148,120 normal images and 97,076 abnormal images. The dataset is available in the source link https://gitee.com/neuhwm/GasHisSDB.git.

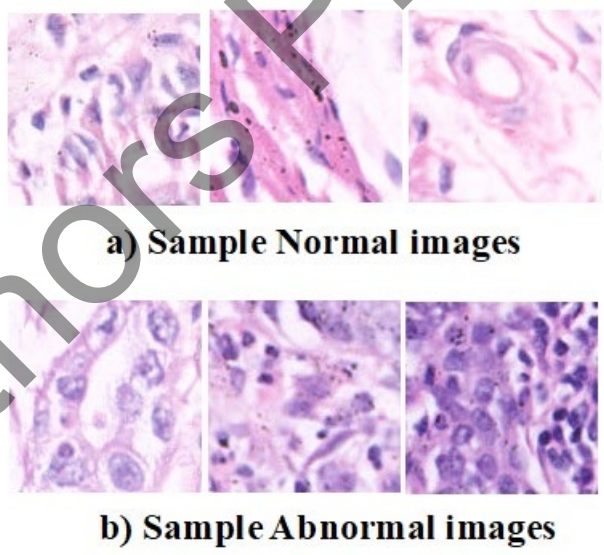


Fig 1. Gastric Cancer Dataset : a) Sample Normal images b) Sample Abnormal images

b) Preprocessing

1.Resize all gastric images to a consistent size

Standardizing the dimensions of gastric images is a critical step in preparing medical datasets for analysis. These images often vary in size and resolution, introducing inconsistencies that can hinder effective model learning. By resizing all images to a uniform dimension, this variability is mitigated, simplifying subsequent processing stages. This resizing process entails adjusting each image to a predetermined width and height, typically determined by computational constraints and

Authors Pre-Proof

the architectural specifications of the chosen convolutional neural network (CNN). For instance, common dimensions for medical image datasets might include 224x224 or 256x256 pixels. Standardization ensures that all images share the same dimensions, facilitating seamless integration into the neural network model for both training and inference tasks.

2.Normalize pixel values to a common scale

In medical imaging, the intensity levels of pixel values can exhibit considerable variation, influenced by factors such as imaging apparatus and configurations. Standardizing these pixel values to a common scale is essential for enhancing model convergence and overall performance. This normalization process entails adjusting the pixel values of each image to conform to a predefined range, typically [0, 1] or [-1, 1]. For instance, a common approach involves dividing all pixel values by 255 to rescale them to the range [0, 1]. By normalizing pixel values, the intensity levels become consistent across all images, facilitating the neural network's ability to discern meaningful patterns and features. This uniformity aids in optimizing model training and enables more accurate predictions based on the standardized representation of image data.

3.Perform data augmentation techniques to increase the diversity of the dataset (e.g., rotation, flipping, scaling)

Data augmentation serves as a prevalent method to expand the size and diversity of the training dataset artificially, achieved by applying transformations to the original images. This strategy is instrumental in mitigating overfitting and bolstering the model's capacity to generalize to unseen data. Various augmentation techniques are employed, including rotation, which alters the image orientation by specific angles such as 90 degrees, flipping to create mirror images horizontally or vertically, aiding in learning invariant features. Scaling resizes images by a certain factor, simulating variations in size, while translation shifts images horizontally or vertically to mimic changes in position. Shearing introduces geometric distortions by applying a shearing transformation to the image. By systematically applying these augmentation techniques to the training dataset, diverse variations of the original images are generated, enriching the dataset's diversity and fortifying the trained model's robustness against unseen data.

c)Improved Inception V3 model for Gastric Cancer Feature Extraction

Figure 2 shows the architecture of the proposed Advancing Gastric Cancer Image Classification model. In the first stage, we employ a pre-trained Inception model, such as Inception-v3 or a variant specifically fine-tuned for gastric cancer image classification, to extract informative features from the gastric cancer images. These features capture high-level patterns and characteristics that are relevant for cancer detection and classification. This model leverages the Inception v3 architecture for feature extraction in gastric cancer classification. This model consists of various components:

1. Base Inception V3

The initial part of the model remains unchanged. It consists of the pre-trained Inception v3 architecture with its convolutional layers. **Freezing Early Layers is described as** the early convolutional layers of Inception v3 are frozen. This preserves the valuable knowledge these layers learned from the pre-trained dataset (like ImageNet) on recognizing basic image features like edges, shapes, and textures. Freezing prevents the model from "forgetting" this foundational knowledge.

2. Feature Extraction

The frozen Inception v3 layers act as a powerful feature extractor. As the image progresses through these layers, it gets transformed into a high-dimensional feature vector. This vector encapsulates the essential information about the image learned by the convolutional layers. The Inception architecture utilizes a series of inception modules that combine convolutional filters of different sizes (1x1, 3x3, 5x5) within a single layer. The convolution operation is defined as follows: Given an input image I of size H × W and a filter/kernel K of size F × F, the filter slides across the input image, computing the dot product of its elements with corresponding elements at each position as shown in equation (1).

This can be represented mathematically as:

Authors Pre-Proof

$$\textbf{Output}(i,j) = \sum_{m=0}^{F-1} \sum_{n=0}^{F-1} I(i+m, j+n) * K(m,n) \quad (1)$$

The stride S determines the step size of the filter movement during convolution. It controls the spatial dimensions of the output feature map. Mathematically, the stride is applied as shown in equation (2).

$$\textbf{Output}(i,j) = \sum_{m=0}^{F-1} \sum_{n=0}^{F-1} I(i*S+m, j*S+n) * K(m,n) \quad (2)$$

Padding is used to prevent shrinking of images during convolution, a technique called padding is used. Padding adds extra pixels, typically zeros, around the edges of the image before applying the convolution filter. This ensures the output maintains the same spatial dimensions as the original image.

3.Inception Modules:

The output of an inception module can be represented as the concatenation of feature maps obtained from different convolutional operations: Output=Concatenate (Conv1,Conv3,Conv5,Conv1x1)

4. New Fully Connected Layers:

- The final fully connected layers of Inception v3 are replaced with new layers specifically designed for gastric cancer classification.
    - **Dense Layer:** A crucial layer is the Dense layer (also called a fully connected layer). This layer takes the feature vector extracted by the previous convolutional layers.
        - **Number of Neurons:** The number of neurons in the Dense layer corresponds to the number of gastric cancer classes you want to predict. For example, if classifying normal, benign, and malignant tissue, the Dense layer would have three neurons (one for each class).
        - **Activation Function:** A softmax activation function is typically used. It converts the output values of the Dense layer into probabilities of belonging to each class.

5. Optional Layers (for Improved Performance):

- While the Dense layer with softmax is essential, consider adding these optional layers depending on your dataset complexity:
    - **Dropout Layer:** To prevent the model from becoming overly reliant on specific features and improve its ability to handle new data, a technique called dropout is used during training. Dropout randomly "switches off" a certain portion of neurons, forcing the network to learn using different combinations of neurons each time. This helps the model generalize better to unseen data and reduces the risk of overfitting.
    - **Additional Dense Layers:** In some cases, using multiple hidden Dense layers with appropriate activation functions (e.g., ReLU) can improve the model's capacity to learn complex, non-linear relationships within the features extracted by Inception v3. However, this increases complexity and requires more training data to avoid overfitting.

The final output of this model is a high-dimensional feature vector for each image. This vector represents the key information about the image learned by the convolutional layers and the new fully connected layers. For feature extraction in gastric cancer classification, an autoencoder for dimensionality reduction is used on the feature vector derived from the improved inception v3. If the feature vector extracted by the new fully connected layers in the Improved Inception v3 model is very high-dimensional, an autoencoder can potentially compress it into a more manageable size for the SVM classifier. This can reduce training time and memory requirements. While Inception v3's features are informative, they can be very high-dimensional (thousands of features) using such high-dimensional data can lead to increased training time for the classifier (e.g., SVM) and increased susceptibility to overfitting, which means the model excels at recognizing training

Authors Pre-Proof

examples but struggles with new data. Dimensionality reduction aims to compress the extracted features into a lower-dimensional representation while retaining the essential information for classification.

d) Autoencoders for Dimensionality Reduction:

- Autoencoder is a special kind of neural network that excels at uncovering hidden patterns in data. They achieve this by learning to represent the data using a smaller set of variables. It consists of two main parts:
    - Encoder: Takes the high-dimensional input features (from Inception v3) and progressively reduces their dimensionality through hidden layers with decreasing numbers of neurons.
    - Decoder: Takes the compressed representation from the encoder (bottleneck layer) and attempts to reconstruct the original features through hidden layers with increasing numbers of neurons.
- During training, the autoencoder learns to:
    - Encode the input features into a lower-dimensional latent space (bottleneck layer). This space captures the most important information for classification based on the autoencoder's learned representation.
    - Reconstruct the original features as accurately as possible using the decoder. This reconstruction loss serves as a training signal, forcing the autoencoder to learn an informative compressed representation.

Benefits of Autoencoder-based Dimensionality Reduction are described by the following Computational Cost is reduced by using the lower-dimensional features from the autoencoder's bottleneck layer can lead to faster training times for the classifier compared to using the high-dimensional features directly from Inception v3. Generalizability is improved by focusing on capturing essential features, autoencoders can potentially help mitigate overfitting, especially with large datasets. Potential for Better Feature Learning is described by Comparing techniques like PCA, autoencoders can learn non-linear relationships between features in the Inception v3 model's output, potentially leading to more informative representations for classification.

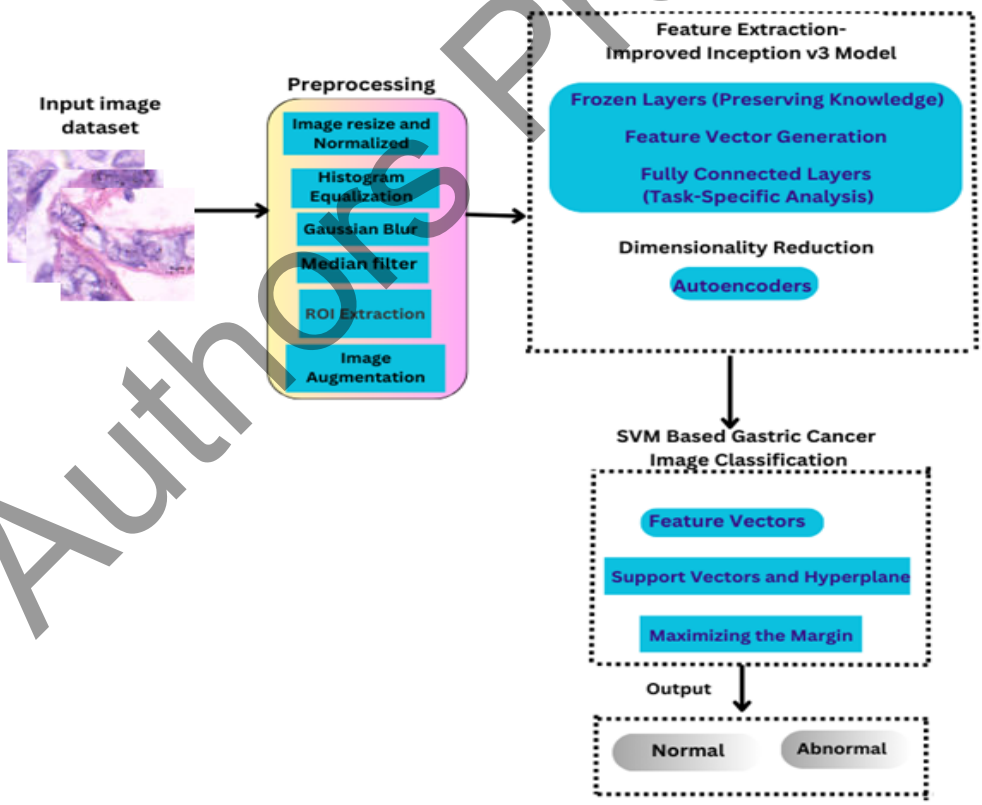


Fig 2 . Architecture of the proposed Advancing Gastric Cancer Image Classification model

Authors Pre-Proof

1. Freeze the early convolutional layers of Inception v3

This retains the valuable knowledge these layers learned from the pre-trained dataset (like ImageNet). These layers are adept at capturing general image features like edges, shapes, and textures. Freezing them prevents the model from "forgetting" this knowledge and allows it to focus on learning new skills. Replacing the final fully connected layers with a Dense layer (with SoftMax activation) is essential for gastric cancer classification. The number of neurons in the Dense layer corresponds to the number of cancer classes we aim to identify. We might also consider adding other layers like dropout or additional Dense layers for improved performance, but this requires careful experimentation and evaluation.

e)SVM Based Gastric Cancer Image Classification

1. Feature Vectors

After processing a gastric cancer image through the Improved Inception v3 model and the autoencoder, a lower-dimensional feature vector representing the image's key characteristics relevant for classification is obtained. This feature vector can be visualized as a point in a multi-dimensional space, where each dimension corresponds to a specific feature extracted (e.g., intensity, texture, edge patterns).

2. Support Vectors and Hyperplane

In the world of high-dimensional feature spaces, SVM acts like a discerning classifier. It searches for the optimal flat separation surface, called a hyperplane, that best divides the data points belonging to different classes. Imagine this hyperplane as a flexible decision boundary that can adapt in higher dimensions beyond the familiar lines and planes of two- and three-dimensional spaces. To achieve the best separation, SVM identifies the most critical data points, the support vectors, acting like gatekeepers on either side of the hyperplane. The margin, the space between the hyperplane and these support vectors, is crucial. A wider margin translates to a clearer separation between the classes, making SVM a powerful tool for tasks like classifying healthy versus cancerous tissue in gastric cancer images.

3. Maximizing the Margin

SVM prioritizes creating a clear distinction between classes. It achieves this by finding a separation boundary, called a hyperplane, that maximizes the distance between itself and the closest data points from each class, known as support vectors. The wider this margin is, the better. Intuitively, a larger margin implies a clearer separation between the classes. This focus on the margin also makes SVM less susceptible to outliers or noisy data in the training set. By prioritizing the most critical data points (support vectors) for the classification boundary, SVM aims to generalize well even on unseen data.

4. Kernel Trick (for Non-Linear Separable Data)

The real world can be messy, and data representing things like cancerous and healthy tissue in gastric cancer images might not always be neatly separable in the initial feature space. This means a straight line (hyperplane) can't perfectly divide the good from the bad. To tackle this challenge, SVM uses a clever method called the kernel trick. This trick acts like a magic portal, projecting the data points into a higher-dimensional space where they become linearly separable. Different types of kernels, like linear kernels (for simpler data) or polynomial and radial basis function (RBF) kernels, can be used for this projection. The best choice of kernel depends on the specific way your data is scattered in its original space.

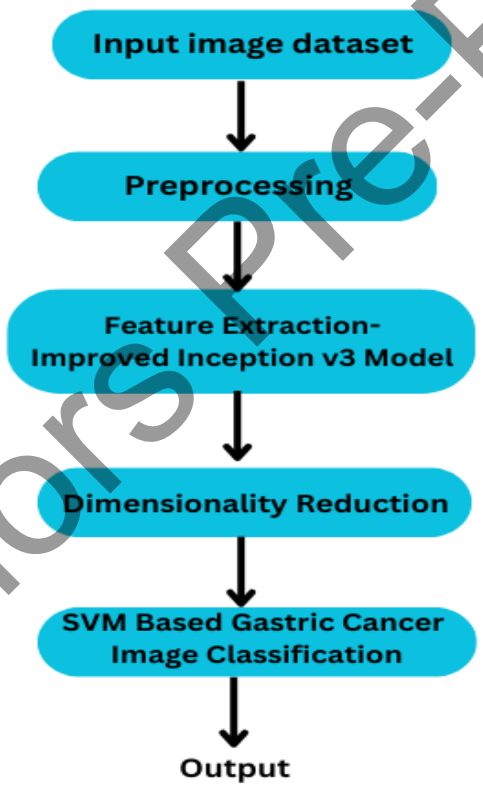
5. Classification of gastric cancer images

Once the SVM is trained on the battlefield of gastric cancer image features, it's ready to face new unseen enemies. The features extracted from a new image are projected into the same high-dimensional space using the same battle plan (kernel function) as during training. This essentially creates a map for the new data point. The trained SVM model then acts as a commander, strategically analyzing the position of the new data point relative to the hyperplane, the decision boundary. Depending on which side of the line the data point falls on, the SVM confidently assigns it to a class, like healthy or cancerous tissue, making a crucial prediction for improved diagnosis.

f. Working principles of the proposed work

Authors Pre-Proof

Working principles of the proposed Enhanced Gastric Cancer Image Classification model is explained in figure 3. The initial phase of the pipeline, data preprocessing, lays the groundwork for successful image analysis. Here, two crucial steps are undertaken: image resizing and pixel value normalization. Resizing ensures all images conform to a standard dimension compatible with the chosen neural network. This prevents issues like excessive padding or information loss during processing. The chosen size balances capturing sufficient detail with computational efficiency. Common sizes for medical image datasets, like 224x224 or 256x256 pixels, are selected based on the specific network and available resources. Normalization, the second step, addresses inconsistencies in pixel intensity caused by variations in imaging devices or settings. By scaling these values to a common range (e.g., 0-1 or -1, 1), the data becomes consistent, allowing the model to learn meaningful patterns from the images more effectively. The second phase, data augmentation, tackles the challenge of limited training data. Here, various techniques are employed to artificially create variations of the original images. This includes rotation, simulating different viewing angles; flipping, mirroring the image horizontally or vertically; scaling, zooming in or out for slight magnification changes; translation, shifting the image to account for potential misalignments; and shearing, introducing minor geometric distortions. Imagine applying these techniques to an image of gastric tissue. The result would be variations that resemble tissue from slightly different viewpoints, magnifications, or with minor positional changes. By effectively increasing the dataset size and diversity through these artificial variations, the model learns robust features. These features allow the model to generalize well to unseen data, ultimately leading to improved classification accuracy when encountering real-world gastric cancer images.



Authors Pre-Proof

Fig 3 . Working principles of the proposed Enhanced Gastric Cancer Image Classification model.

The third phase dives into feature extraction, a crucial step in image analysis. Here, the pipeline utilizes an improved version of the Inception v3 model. This pre-trained model is a powerhouse, having already learned to recognize fundamental image features like edges, shapes, and textures from a massive dataset like ImageNet. These learned capabilities are directly applicable to the task of gastric cancer image classification. To leverage this pre-trained knowledge effectively, the early layers of Inception v3 are frozen. This preserves the valuable information about generic image features learned from the massive dataset. Meanwhile, newly added layers are tasked with analyzing the processed images. The Inception v3 model, with both frozen and new layers, transforms each gastric cancer image into a high-dimensional feature vector. This vector acts like a fingerprint, capturing the essential details and key characteristics relevant for classifying the image as healthy or cancerous tissue. The newly added layers then specifically analyze these feature vectors to identify

patterns that differentiate between the two classes. The fourth phase tackles the challenge of high-dimensional feature vectors generated by Inception v3. Here, an autoencoder comes into play. This powerful tool consists of two parts: an encoder and a decoder. The encoder takes the complex feature vectors from Inception v3 and progressively reduces their dimensionality by passing them through hidden layers. Imagine squeezing a complex ball of yarn into a tighter ball. This compressed version, called the bottleneck layer, retains the most critical information for classification while being much more efficient to handle in the next stage. The decoder's role is fascinating. It receives this compressed representation and attempts to reconstruct the original feature vectors. While the reconstruction might not be perfect, the process itself helps the autoencoder refine and solidify its understanding of the essential details captured within the feature vectors. This dimensionality reduction allows for faster and more manageable processing in the final classification stage, without sacrificing the crucial information needed to distinguish healthy from cancerous tissue.

IV. RESULTS AND ANALYSIS

This section presents the performance metrics used in this work and the results obtained from the proposed model, along with a comparative analysis.

IV.i) Performance Metrics

The assessment of classification algorithms' effectiveness relies on the accuracy parameter. However, when applied to evaluate a model trained on imbalanced data, achieving accuracy becomes challenging and may result in a performance matrix that is misleading, as depicted in equation (3).

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} \quad (\mathbf{3})$$

Recall serves as a dedicated metric to gauge the efficiency of a classifier. Equation (4) illustrates recall, which is the measure of correct classifications achieved by the classification model.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \quad (\mathbf{4})$$

Precision, as indicated in Equation (5), is formulated by dividing the true positive instances by the sum of the true positive and false positive (FP) instances.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \quad (\mathbf{5})$$

The Area under the Curve (AUC) summarizes the ROC curve, offering a unified assessment of a binary classification model's overall performance. It quantifies the likelihood that a positive sample selected at random will receive a higher classification score than a negative sample, based on the model's predictions. The formula for AUC can be calculated using numerical integration techniques or the trapezoidal rule. Using the trapezoidal rule, the formula for AUC is shown in equation (6).

$$\text{AUC} = \frac{\sum[(\text{i} = 1) \text{ to } (\text{n} - 1)] \, (\text{TPR[i]} + \text{TPR[i} + 1]) * (\text{FPR[i} + 1] - \text{FPR[i]})}{\mathbf{2}} \quad (\mathbf{6})$$

Where TPR[i] and FPR[i] represent the TPR and FPR values at the i^{th} point on the ROC curve, and n is the total number of points on the curve.

IV.ii) Results

Following feature extraction using the improved InceptionV3 model, the sample feature vector set is represented as shown in the figure 4

Authors Pre-Proof

```
Concatenated feature vectors shape: (10, 131072)
Concatenated feature vectors:
[[0.         0.         0.         ... 0.1905391  0.         0.7115126 ]
 [0.         0.03634824 0.4111221  ... 0.05402149 0.         1.1440314 ]
 [0.         0.8561381  0.         ... 0.         0.757122   0.9653863 ]
 ...
 [0.         0.         0.         ... 0.6365009  2.9426198  0.7865159 ]
 [1.2266221  0.         0.5278355  ... 1.1687375  0.         0.        ]
 [0.         0.         0.62589467 ... 0.         0.         0.58563757]]
```

Fig 4. Sample Feature Vector after Feature Extraction

After feature extraction, an autoencoder is used to reduce the dimensionality of the features. The figure 5 illustrates the output of the encode feature vector.

```
Shape of encoded feature vector: (1, 131072)
Encoded feature vector:
[[1.5527715e-21 2.7152565e-22 6.6352975e-01 ... 3.7410657e-22
  2.8565052e-22 6.1686379e-01]]
```

Fig 5. Feature Vector after dimensionality reduction

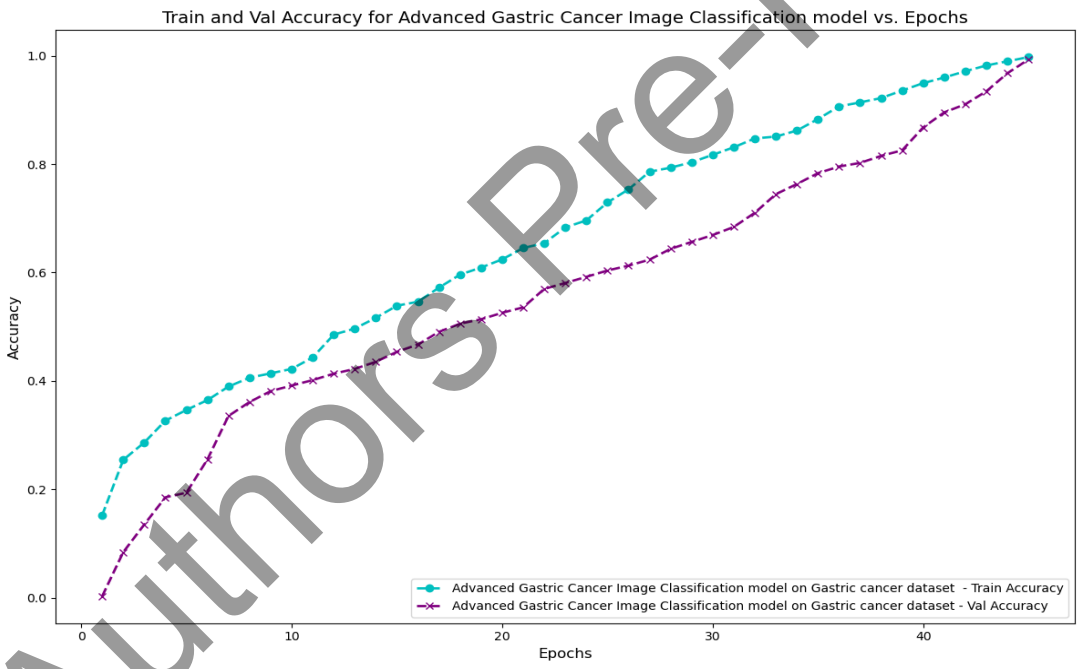


Fig 6. Advancing Gastric Cancer Image Classification model's training, valid accuracy on the GasHisSDB for Epochs 45

This figure 6 shows the training and validation accuracy of a machine learning model for classifying advanced gastric cancer images. The x-axis is labelled "Epochs," and the y-axis is labelled accuracy." There are two lines on the graph, one for "Train Accuracy" and another for "Validation Accuracy". The model's training accuracy increases as the number of epochs increases. This suggests that the model is learning to classify the images correctly based on the training data. The validation accuracy also increases as the number of epochs increases. This suggests that the model is generalizing well and is not simply overfitting the training data. The model performed well on a test set; this would give a better indication of how well the model would perform on unseen data. Starting at 0.15, the training accuracy increases as the epoch increases. The highest training accuracy achieved is 0.9976. The validation accuracy also varies across epochs, with the lowest being 0.052 and the highest being 0.9932.

Authors Pre-Proof

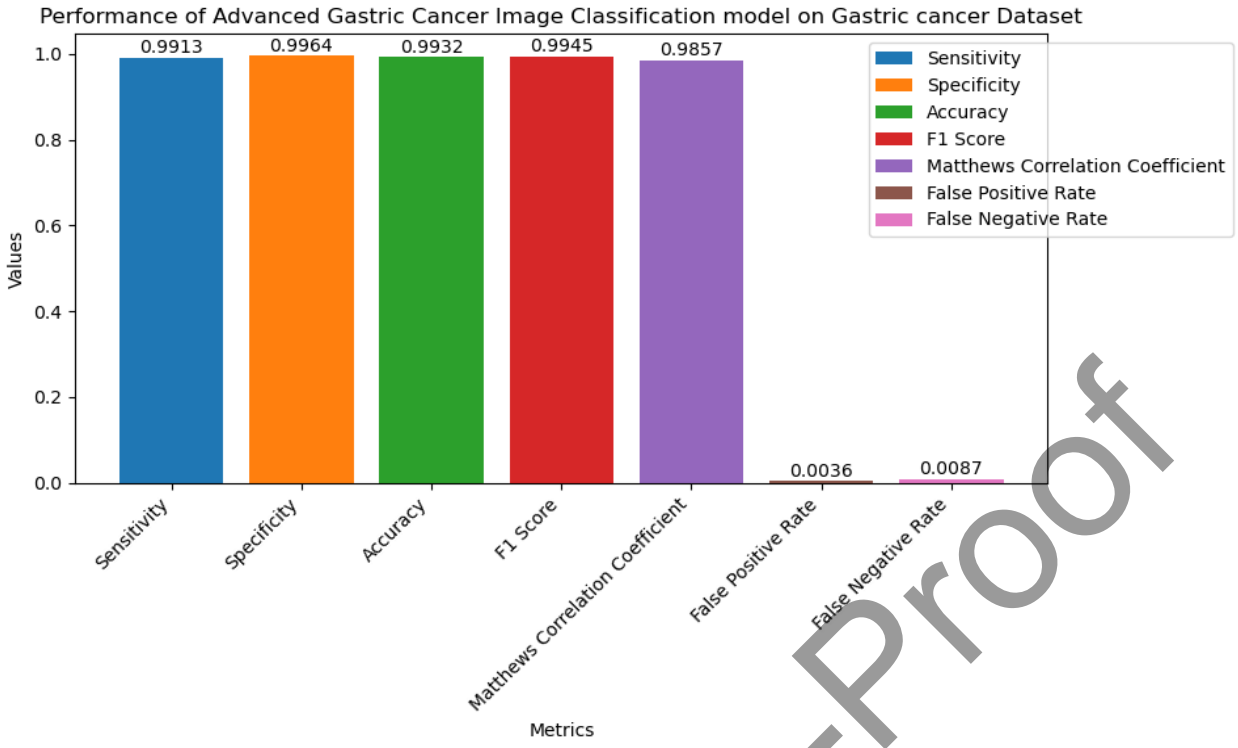


Fig 7. Performance of the proposed Enhanced Gastric Cancer Image Classification model

This figure 7 illustrates the performance of a model on a classification task. The model achieved a high sensitivity of 0.9913, indicating it correctly classified 99.13% of actual positive cases. Similarly, the specificity of 0.9964 signifies that the model accurately classified 99.64% of actual negative cases. Examining errors, the model produced a low false positive rate of 0.0036, meaning only 0.36% of actual negative cases were incorrectly classified as positive. There was also a low false negative rate of 0.0087, indicating that the model misclassified just 0.87% of actual positive cases as negative. Overall, the model achieved a high accuracy of 0.9932, signifying it correctly classified 99.32% of all cases.

IV.iii) Analysis

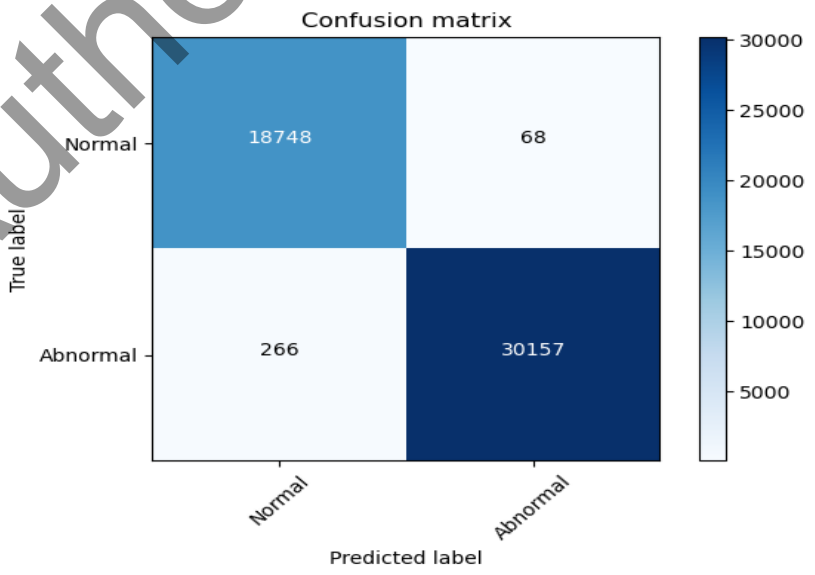


Fig 8 . Confusion matrix of the proposed gastric cancer classification model.

Authors Pre-Proof

This figure 8 illustrates the confusion matrix that evaluates a medical image classification model's performance in distinguishing between normal and abnormal images. Here, 'normal' signifies images free from disease, while 'abnormal' represents those containing the disease. True positives (TP), at 18,748, represent abnormal images correctly classified. False positives (FP), at 68, indicate normal images incorrectly classified as abnormal. Conversely, false negatives (FN), at 266, represent abnormal images missed by the model and classified as normal. Finally, true negatives (TN), at 30,157, represent normal images correctly identified.

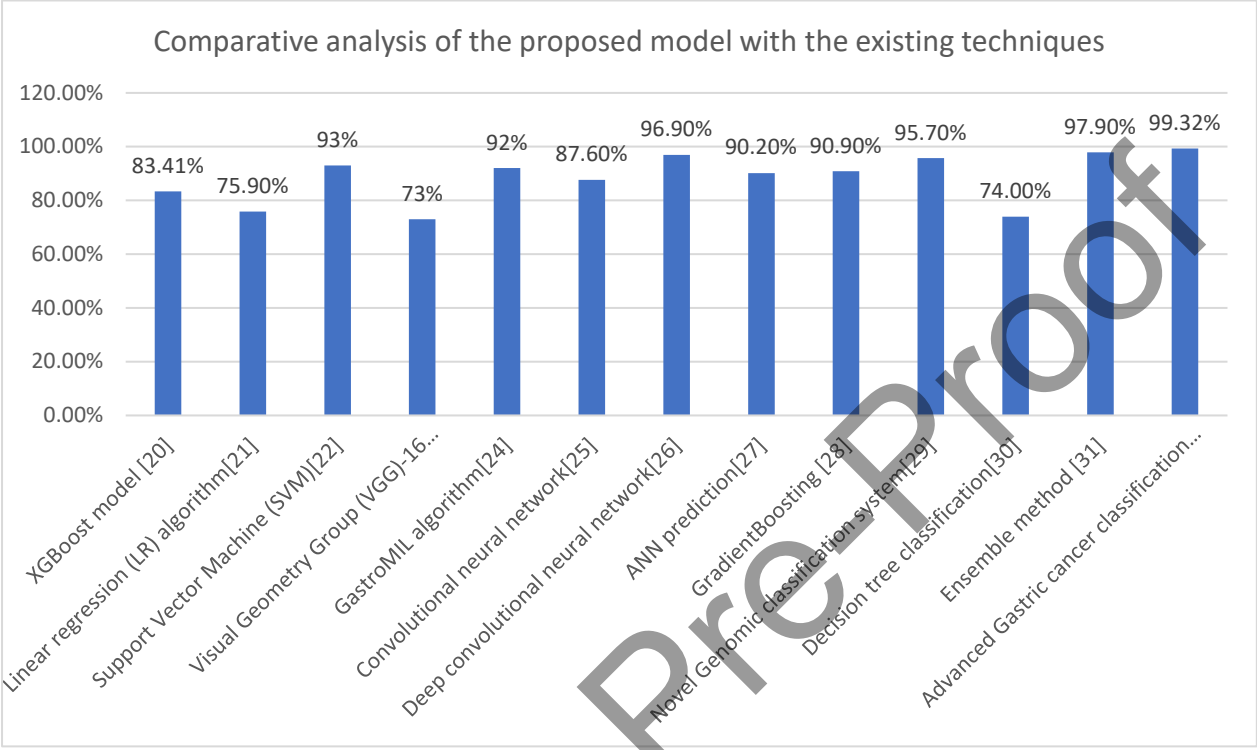


Fig 9. Accuracy based comparison of the proposed gastric cancer classification model with the existing techniques.

Figure 9 compares the performance of various classification algorithms on a gastric cancer classification task. The X-axis lists various classification algorithms including XGBoost model, Linear Regression (LR), Support Vector Machine (SVM), Visual Geometry Group (VGG)-16, GastroMIL algorithm, Convolutional Neural Network (CNN), Deep Convolutional Neural Network (DCNN), ANN prediction, Gradient Boosting, Novel Genomic classification, Decision Tree classification, Ensemble method, and the proposed Advanced Gastric cancer classification model. The Y-axis represents the percentage accuracy achieved by each algorithm on the classification task. 100% represents perfect classification, while 0% indicates completely random assignment.The bars show that the proposed model, likely a Deep Convolutional Neural Network (DCNN) based on its location on the X-axis, achieves the highest accuracy (around 99.32%) compared to other algorithms. The performance of the remaining algorithms varies, with some exceeding 90% accuracy (e.g., XGBoost model at 97.90%) and others falling below 80% (e.g., Linear Regression at 75.90%).

V. CONCLUSION AND FUTURE WORK

This research presented a deep learning-based approach for gastric cancer classification using image analysis. The proposed method addressed limitations of existing techniques by employing deep feature extraction, dimensionality reduction, and classification strategies. This approach achieved high accuracy (99.32%), sensitivity (99.13%), and specificity (99.64%) in classifying gastric cancer. These results suggest that the proposed method has the potential to be a valuable tool for automated gastric cancer diagnosis, potentially improving efficiency and accessibility compared to traditional methods. Future work could explore the generalizability of this approach on larger and more diverse datasets, paving the way for its real-world implementation in clinical settings. And this work focuses on single-modality features. Multi-modality feature fusion could be investigated in future work to potentially improve performance.

Authors Pre-Proof

References

[1]. Li Y, Deng L, Yang X, Liu Z, Zhao X, Huang F, Zhu S, Chen X, Chen Z, Zhang W. Early diagnosis of gastric cancer based on deep learning combined with the spectral-spatial classification method. Biomedical optics express. 2019 Oct 1;10(10):4999-5014.

[2]. Hu Y, Bao H, Jin H, Zhao J, Xu Y, Huang X, Liu S, Lu B. Performance evaluation of four prediction models for risk stratification in gastric cancer screening among a high-risk population in China. Gastric cancer. 2021 Nov;24(6):1194-202.

[3]. Martin-Romano P, Sola JJ, Diaz-Gonzalez JA, Chopitea A, Iragorri Y, Martínez-Regueira F, Ponz-Sarvise M, Arbea L, Subtil JC, Cano D, Ceniceros L. Role of histological regression grade after two neoadjuvant approaches with or without radiotherapy in locally advanced gastric cancer. British Journal of Cancer. 2016 Sep;115(6):655-63.

[4]. Correa P, Piazuelo MB. The gastric cancer. Colombia Médica. 2013 Sep 30;44(3):192-201.

[5]. Hou JZ, Dong SS, Yuan M, Zhong C. Patterns of death and life lost of gastric cancer in China cancer registration areas, 2013. Chinese Journal of Cancer Prevention and Treatment 2019; 26:986–990.

[6]. Vetrithangam, D., Senthilkumar, V., Neha, A., Naresh, P., & Kumar, M. S. (2022). Coronary artery disease prediction based on optimal feature selection using improved artificial neural network with meta-heuristic algorithm. Journal of Theoretical and Applied Information Technology, 100(24).

[7]. Mahmud MS, Huang JZ, Fu X. Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. International Journal of Computational Intelligence and Applications. 2020 Mar 13;19(01):2050002.

[8]. Kantamaneni, p., Vetrithangam, d., Saisree, m. m., Shargunam, s., Kumar, s. s., & Bekkanti, a. (2023). Optimized fuzzy c-means (fcm) clustering for high-precision brain image segmentation and diagnosis using densenet features. Journal of Theoretical and Applied Information Technology, 101(24).

[9]. Khan MA, Majid A, Hussain N, Alhaisoni M, Zhang YD, Kadry S, Nam Y. Multiclass stomach diseases classification using deep learning features optimization.

[10]. Vetrithangam, D., Pegada, N. K., Himabindu, R., & Kumar, A. R. (2024, March). A state of art review on image analysis techniques, datasets and applications. In *AIP Conference Proceedings* (Vol. 3072, No. 1). AIP Publishing.

[11]. Guan X, Lu N, Zhang J. Computed tomography-based deep learning nomogram can accurately predict lymph node metastasis in gastric cancer. Digestive Diseases and Sciences. 2023 Apr;68(4):1473-81.

[12]. Vetrithangam, d., shruti, p., arunadevi, b., himabindu, r., kumar, p. N., kumar, a. R., ... & arnet zitha, d. R. (2023). Optimum feature selection based breast cancer prediction using modified logistic regression model. Journal of theoretical and applied information technology, 101(8).

[13]. Vetrithangam, D., arunadevi, b., kumar, a. K., & nalini, s. (2023). Olgv3 net: optimized lightgbm with inceptionv3 for accurate multi-class breast cancer image classification. Journal of theoretical and applied information technology, 101(24).

[14]. Wang Z, Liu Y, Niu X. Application of artificial intelligence for improving early detection and prediction of therapeutic outcomes for gastric cancer in the era of precision oncology. InSeminars in Cancer Biology 2023 Apr 26. Academic Press.

[15]. Niu PH, Zhao LL, Wu HL, Zhao DB, Chen YT. Artificial intelligence in gastric cancer: Application and future perspectives. World journal of gastroenterology. 2020 Sep 9;26(36):5408.

[16]. Chae JW, Cho HC. Enhanced classification of gastric lesions and early gastric cancer diagnosis in gastroscopy using multi-filter AutoAugment. IEEE Access. 2023 Mar 23;11:29391-9.

[17]. Ma L, Lu G, Wang D, Wang X, Chen ZG, Muller S, Chen A, Fei B. Deep learning based classification for head and neck cancer detection with hyperspectral imaging in an animal model. InMedical imaging 2017: Biomedical applications in molecular, structural, and functional imaging 2017 Mar 13 (Vol. 10137, pp. 632-639). SPIE.

[18]. Yoon HJ, Kim S, Kim JH, Keum JS, Oh SI, Jo J, Chun J, Youn YH, Park H, Kwon IG, Choi SH. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. Journal of clinical medicine. 2019 Aug 26;8(9):1310.

[19]. Yong MP, Hum YC, Lai KW, Lee YL, Goh CH, Yap WS, Tee YK. Histopathological gastric cancer detection on GasHisSDB dataset using deep ensemble learning. Diagnostics. 2023 May 18;13(10):1793.

[20]. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC gastroenterology. 2023 Jan 10;23(1):6.

[21]. Zhou CM, Wang Y, Yang JJ, Zhu Y. Predicting postoperative gastric cancer prognosis based on inflammatory factors and machine learning technology. BMC Medical Informatics and Decision Making. 2023 Mar 31;23(1):53.

[22]. Talebi A, Celis-Morales CA, Borumandnia N, Abbasi S, Pourhoseingholi MA, Akbari A, Yousefi J. Predicting metastasis in gastric cancer patients: machine learning-based approaches. Scientific Reports. 2023 Mar 13;13(1):4163.

[23]. Yoon HJ, Kim S, Kim JH, Keum JS, Oh SI, Jo J, Chun J, Youn YH, Park H, Kwon IG, Choi SH. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. Journal of clinical medicine. 2019 Aug 26;8(9):1310.

[24]. Huang B, Tian S, Zhan N, Ma J, Huang Z, Zhang C, Zhang H, Ming F, Liao F, Ji M, Zhang J. Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: A retrospective multicentre study. EBioMedicine. 2021 Nov 1;73.

[25]. Sakai Y, Takemoto S, Hori K, Nishimura M, Ikematsu H, Yano T, Yokota H. Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. In2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2018 Jul 18 (pp. 4138-4141). IEEE.

[26]. Zheng X, Wang R, Zhang X, Sun Y, Zhang H, Zhao Z, Zheng Y, Luo J, Zhang J, Wu H, Huang D. A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology. Nature communications. 2022 May 19;13(1):2790.

[27]. Nilsaz-Dezfouli H, Abu-Bakar MR, Arasan J, Adam MB, Pourhoseingholi MA. Improving gastric cancer outcome prediction using single time-point artificial neural network models. Cancer informatics. 2017 Feb 14;16:1176935116686062.

[28]. Zhou C, Wang Y, Ji MH, Tong J, Yang JJ, Xia H. Predicting peritoneal metastasis of gastric cancer patients based on machine learning. Cancer Control. 2020 Oct 27;27(1):1073274820968900.

[29]. Wang H, Ding Y, Chen Y, Jiang J, Chen Y, Lu J, Kong M, Mo F, Huang Y, Zhao W, Fang P. A novel genomic classification system of gastric cancer via integrating multidimensional genomic characteristics. Gastric Cancer. 2021 Nov;24:1227-41.

[30]. Mohammadzadeh F, Noorkojuri H, Pourhoseingholi MA, Saadat S, Baghestani AR. Predicting the probability of mortality of gastric cancer patients using decision tree. Irish Journal of Medical Science (1971-). 2015 Jun;184:277-84.

[31]. Prince MS, Hasan A, Shah FM. An efficient ensemble method for cancer detection. In2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) 2019 May 3 (pp. 1-6). IEEE.

Authors Pre-Proof