IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Advancing Cancer Classification with Hybrid Deep Learning: Image Analysis for Lung and Colon Cancer Detection

[1]Abdus Sobur, [2]Md Imran Chowdhury Rana, [3]Md Zakir Hossain, [4]Anwar Hossain, [5]Md Firoz Kabir
[1]Graduate Student, [2]Bachelor Student, [3]Graduate Student, [4]Graduate Student, [5]Graduate Student
[1] Masters of Information Technology
Westcliff University, USA
[2]Department of Bachelor of Business Administration
International American University ,USA
[3]Master's in Data Science
Grand Canyon University, USA
[4]Mastersof Information Science and Technology
California State University,USA
[5]Masters in Information Technology
University of the Cumberlands, USA

Abstract:This paper presents a groundbreaking approach to classifying lung and colon cancer from high-dimensional histopathological images, employing an advanced hybrid deep learning model. Our dataset encompasses 20,000 high-resolution images across five distinct classes, posing significant challenges in terms of computational efficiency and model accuracy. The dual-CNN structure of our model ensures a comprehensive extraction of both fine-grained and global features, crucial for accurate cancer classification. Through innovative feature reduction techniques, we effectively mitigate the curse of dimensionality, enhancing the model's computational efficiency and robustness. The integration of global average pooling and dense layers with dropout regularization further contributes to the model's performance, preventing overfitting and ensuring generalizability. Our approach achieves an impressive classification accuracy of 99%, demonstrating the model's capability to handle high-dimensional datasets with precision. This work marks a significant contribution to medical image analysis, providing a reliable and efficient solution for cancer classification and setting a new standard in the field.

***Index Terms* -** Feature reduction,DL,ML,PCA,Rationale,Feature,F1-Score,Recall,Accuracy.

I. INTRODUCTION

In the contemporary landscape of medical research and diagnostics, the utilization of Ml, particularly DL, has emerged as a transformative force, revolutionizing the way we approach, analyze, and interpret medical data. Among the myriad of applications, the classification of cancerous tissues through histopathological images stands out as a critical area of focus, with colon cancer being a predominant subject of study due to its high incidence and mortality rates worldwide. The early detection and accurate classification of colon cancer are imperative for effective treatment planning and improving patient outcomes, making the role of advanced computational models in medical imaging more crucial than ever.However, the journey towards achieving high accuracy in cancer classification is fraught with challenges, especially when dealing with high-dimensional datasets. Histopathological images are inherently high-resolution, resulting in datasets with a large number of features. While this high dimensionality holds valuable information for accurate classification, it also introduces the 'curse of dimensionality', a phenomenon where the performance of a machine learning model degrades as the number of features increases, due to the sparsity of the data in highdimensional space. This not only makes the training of models computationally intensive but also poses risks of overfitting, where the model becomes too tailored to the training data, losing its ability to generalize well to new, unseen data.Addressing these challenges necessitates innovative approaches in model architecture and data processing. Feature reduction methods have emerged as a powerful solution, aiming to reduce the dimensionality of the dataset while preserving its informative content. By doing so, these methods enhance the computational efficiency of the model, mitigate the risks of overfitting, and often lead to improved model performance. However, the application of feature reduction methods in the context of high-dimensional histopathological images for colon cancer classification is a complex task, requiring a careful balance between dimensionality reduction and the preservation of crucial information for accurate classification.In this paper, we introduce an advanced hybrid deep learning model, specifically designed to navigate the complexities of high-dimensional histopathological image datasets for colon cancer classification. Our approach is grounded in a dual-CNN structure, ensuring a comprehensive extraction of features, followed by innovative feature reduction techniques to enhance the model's efficiency and performance. This work is not just a technical endeavor but also a contribution to the field of medical research, aiming to enhance our capabilities in early cancer detection and classification, ultimately contributing to better patient care and outcomes.

II. RELATED WORK

Guney and Oztoprak (2022)**[1]** carried out a ground-breaking study to use ensemble feature selection (EFS) to increase the robustness and accuracy of machine learning models. By proposing the Minimum Weight Threshold Method-based EFS (MWTEFS), they solved the problems brought up by outliers in ranked feature lists. By using the Support Vector Classifier for weight assignment and the MWT technique for handling outliers, they produced a robust framework. Their findings showed how well the approach reduced computer complexity, maintained classification efficiency, and improved the stability of gene selection.

Ebrahimpour et al. (2017)**[2]** investigated the complexities of high-dimensional microarray datasets, which are notorious for having small sample sizes and an excessive number of characteristics.The significance of feature selection in optimising the efficacy of models created with these kinds of datasets was stressed. To get around the problems, they proposed a two-stage feature selection procedure that coupled the Occam's Razor idea with the reduced row Echelon form. Finding linearly independent traits was the aim of this innovative approach, which would ensure a more precise and effective model performance.

Sabzi, S. Abbaspour Gilandeh, Y., and Garcia Mateos, F. (2017)**[3]** used a dataset of 300 colour photos from three common types to creatively handle the problem of reliably recognising distinct orange varieties. From each image, they extracted a full set of 263 parameters, which they then used to pick features using a hybrid ANN-PSO technique. As a result, the six most significant features were determined, which prepared the ground for a comparison of three classifiers: ANN-ABC, ANN-HS, and kNN. The ANN-ABC model improved the results even more with an accuracy rate of 96.70%, while the hybrid ANN-HS model demonstrated remarkable performance with an astounding 94.28% accuracy. These results considerably outperformed the 70.9% accuracy of the kNN model, indicating the effectiveness of the suggested methodology and its potential for broad implementation in fruit variety categorization inside processing factorie

In order to overcome the difficulties in Hyperspectral Image Classification (HSIC), Ahmad et al. (2021)**[4]** proposed a compact hybrid CNN model for enhanced feature extraction. Their method outperforms 2D CNNs alone while reducing the computational cost of 3D CNNs by balancing spatial-spectral feature extraction between 2D and 3D layers. A lot of preprocessing was done, including different dimension reduction techniques, to improve the classification results and shorten the computing time. With the exception of a few computationally demanding alternatives, the suggested model outperformed state-of-the-art CNN models in terms of statistical significance and generalisation performance. By providing a well-rounded and effective method for feature extraction and categorization, this study represents a substantial development in HSIC.

An empirical examination of the effects of feature reduction in deep learning and traditional approaches, specifically for foot image classification in knee rehabilitation, was carried out by Jaruenpunyasak and Duangsoithong (2021)**[5].** Their goal was to reduce memory and computing expenses for low-power devices by using convolutional and dense autoencoders for deep learningbased feature reduction. The outcomes were compared to more traditional techniques, like local binary pattern algorithms and histograms of directed gradients, employing classifiers like support vector machines, k-nearest neighbours, and multilayer perceptrons. The study concluded that because conventional methods could project pixels onto a histogram, they were able to obtain higher accuracy with fewer characteristics. Moreover, it was found that deep learning layers retained a high degree of accuracy even with a reduction in features, suggesting the possibility of effective edge computing applications.

In 2018, El-Dahshan and Bassiouni **[6]** introduced a thorough approach to automatically classify MR pictures of the human brain. The procedure consists of five basic steps: MR image noise reduction, stationary wavelet transformation for feature extraction, component analysis and KLDA for feature reduction, and classification. The data is acquired from Harvard and an Egyptian database. To classify the MRI pictures as normal or pathological, two classifiers were developed: Levenberg-Marquardt (LM-ANN) and K-Nearest Neighbour (KNN) on Euclidean distance. In contrast to previous recent efforts in the field, both classifiers remarkably achieved 100% classification accuracy, demonstrating the robustness and efficacy of the suggested approaches and setting a high bar.

Significant progress was made in the application of intelligent computing for environmental analysis by Ghosh et al. (2023) when they developed a method for evaluating water quality**[7]** through the use of predictive machine learning.Rahat (2023)**[8]** examined the complexity of these tumours using deep learning for Flair segmentation and genetic analysis of brain MR images in order to better understand and diagnose lower-grade gliomas.Ghosh (2023)**[9]** employed convolutional neural networks to detect and forecast potato leaf diseases, underscoring the potential of deep learning in the management of crop diseases.Mandava (2023)**[10]** provided an inclusive approach that combines machine learning and deep learning to predict cardiovascular illness in the Bangladeshi population, indicating a promising direction for healthcare analytics.

EBM3GP, a novel Evolutionary Bi-objective Genetic Programming-based method for dimensionality reduction in hyperspectral image (HSI) classification, was introduced by Zhou et al. (2023)**[11]**. By extracting both high-level and low-level (bands) features simultaneously from raw HSI spectra, this unsupervised method overcomes the drawbacks of single-level feature extraction techniques. With the use of multi-dimensional trees for encoding, two mutually limiting metrics for evaluation in the absence of HSI pixel labels, and population evolution to optimise multiple trees, EBM3GP produces a Pareto optimal individual that is decoded as a DR strategy. The method proved to be effective and reliable, particularly for small-scale HSI datasets, as it outperformed five widely-used DR techniques on three HSI datasets. In order to improve diagnosis accuracy and efficiency in healthcare, Ghosh, Rahat, Mohanty, Ravindra, and Sobur (2024)**[12]** did a study on ML and DL approaches for skin cancer detection.

To help in the early diagnosis and treatment of diabetic retinopathy, Kundu et al. (2022)**[13]** created a hierarchical U-Net based framework for the segmentation of red lesions in retinal fundus pictures. To reduce false positives, the framework incorporates a sub-image classification technique, with different sub-image sizes assessed for best results. Because semantic features and fine details could be captured by the stacked U-Net design, there were fewer false negatives and fewer false positives overall. With better sensitivity, accuracy, and F1-Score than cutting-edge networks like U-Net and attention U-Net, the suggested technique outperformed them on the DIARETDB1 dataset, demonstrating its promise as an automated screening tool for diabetic retinopathy.

Mandava et al. (2023) demonstrated the use of advanced analytics in the fight against agricultural diseases by using DL approaches to identify and classify yellow rust**[14]** infection in wheat.Khasim (2023)**[15]** discussed the application of deep and machine learning for real-time identification and diagnoses of rice-leaf disorders in Bangladesh, highlighting the significance of technology on agricultural health.

Ghosh (2023)**[16]** examined the application of deep learning and machine learning in the intelligent photo recognition of microorganisms as part of their examination of the issues and advancements in the field.Mohanty, Ghosh, Rahat, and Reddy (2023)**[17]** investigated state-of-the-art deep learning models for the classification of maize leaf illnesses in a field study conducted in Bangladesh, enhancing crop disease identification.

In 2022, Chen **[18]** presented a new method for classifying sketches images, solving problems with the approaches' inadequate convergence and the expanding needs of network applications. The study performed classification studies on the semantic aspects of sketch images using the concepts of deep learning. Convolutional neural networks, convolution feature models, and the extraction of sketch boundaries were all thoroughly examined in this process. The studies' findings showed that the suggested convolution classification and recognition method worked better than conventional classification methods. It proved more proficient at satisfying the objectives of network intelligent processing for sketch image feature classification, exhibiting improved accuracy in dimensionality reduction and error rate detection.

A technique for obtaining variable spectral channels to improve hyperspectral image classification was proposed by Serpico and Moser **[19]** in 2007. In order to improve classification accuracy, they created a process to average contiguous channels of the hyperspectral image, resulting in spectral bands. Three search algorithms that were derived from feature-selection approaches and an interclass distance on a training set were used to tackle the optimisation problem. The approach worked well, outperforming previous well-liked methods and maintaining the new spectral bands' interpretability.

The susceptibility of Deep Learning models to noise and redundant information in high-dimensional picture datasets was examined by Pintelas et al. (2021)**[20].** In order to minimise dimensionality and filter out noise, they suggested using a convolutional autoencoder topology to provide robust feature representations. Convolutional neural networks receive the compressed output, which significantly improves their performance. This method shows how well DL and unsupervised dimensionality reduction work together to produce reliable image classification.

Wang et al. (2018)**[21]** investigated how high-dimensional characteristics affected the classification of hyperspectral images while recognising the difficulties associated with training, computing, and storing them. They demonstrated the multi-local binary pattern descriptor's superiority over conventional and previously suggested techniques by introducing a texture feature based on it. They used SVM for classification and PCA for dimension reduction to make high-dimensional features feasible. Experiments on two genuine hyperspectral image datasets showed that high-dimensional features were more accurate than their lowerdimensional counterparts.

The problem of overfitting in the context of dimensionality reduction for high-dimensional picture data categorization was examined by Liu and Gillies (2016)**[22].** They refuted the widely held notion that increasing inter-class discrimination improves classification performance, showing that in high-dimensional datasets, this strategy might actually cause significant overfitting. Their theoretical research supported a decrease in inter-class discrimination by confirming the existence of perfectly discriminative subspace projections. They presented a novel dimensionality reduction methodology called "Soft Discriminant Maps," consistently outperforming existing approaches in classification performance and demonstrating a clear correlation between the degree of inter-class discrimination and the effectiveness of the classification.

Two tensor techniques, M3PSTM and M3VSTM, were proposed by Zhang and Chow (2012)**[23]** for picture segmentation and classification. These techniques guarantee computational performance in high-dimensional datasets by representing images as matrices. They make use of the maximum margin criterion for tensor space robust classification. Their effectiveness was validated by extensive experiments on a variety of datasets, which demonstrated either superior or equivalent performance to existing methods.

A thorough methodology for the segmentation and classification of auto parts was described by Lin, Yu, and Chen (2022)**[24],** making use of both 2D texture information and 3D spatial data. The work integrates 2D image processing and 3D model analysis to provide automatic marking algorithms for automotive part detection and categorization. Three processing phases are applied to the 2D system:automated segmentation, background-variable picture generation, and YOLOv3 model-based car part identification. The two-stage processed 3D model uses a trained PointNet model for part recognition on 3D car models and combines a 3D triangular grid with texture pictures for part identification. In both 2D and 3D studies, the approach showed excellent precision and accuracy, demonstrating its potential for use in autonomous car systems and other technology sectors.

Amin, Kabir, and Shobur in 2023 The purpose**[25]** of this project is to improve operational efficiencies and business decisions by analysing Walmart's data using machine learning techniques. The study demonstrates the potential of machine learning in retail management and customer support enhancement.Sobur, Kabir, Islam, 2023 In order to protect**[26]** children in the cyberspace, this study highlights the serious effects of cyberbullying on children's rights and the necessity of digital education and legal frameworks.Kabir, Amin, Sobur, 2023 The authors**[27]** create a machine learning model to forecast stock prices, showcasing its precision and potential utility as a tool for financial analysts and investors conducting market research.

The study by Shobur et al. in the International Journal of Creative Research Thoughts compares physical and cyberspace social engineering attacks, highlighting defense mechanisms against these threats. It provides insights into the evolving landscape of social engineering tactics and emphasizes the necessity for robust security strategies to mitigate such attacks effectively**[28]**.Shobur and associates, 2023 This paper**[29]** presents a thorough overview of defence techniques against these increasingly sophisticated threats by drawing comparisons between physical and cyberspace social engineering attacks**[Table.1].**

Table.1 Summary of the Related Work

Author(s)	Year	Title	Main Contribution	Methodology	Outcome
[Guney et al]	2022	Ensemble Feature Selection (EFS)	Proposed MWT-EFS for robust machine learning models.	Support Vector Classifier, MWT for outlier handling.	Improved robustness, reduced complexity, maintained efficiency.
[Ebrahimpour et al]	2017	Two-stage feature selection	Addressed highdimensional microarray datasets' challenges.	Occam's Razor, reduced row Echelon form.	Precise, effective model performance with independent traits.
[Sabzi et al]	2017	Feature selection for orange variety recognition	Utilized hybrid ANN-PSO technique for feature selection.	ANN-ABC, ANN-HS, kNN classifiers.	High accuracy in fruit variety categorization.
[Ahmad et al]	2021	Hybrid CNN model for HSIC	Enhanced feature extraction for HSIC.	Hybrid 2D/3D CNN model, preprocessing, dimension reduction.	Outperformed state-of-the-art CNNs in generalisation performance.
[Jaruenpunyasak et al]	2021	Feature reduction in foot image classification	Empirical examination of feature reduction techniques.	Convolutional, dense autoencoders, traditional techniques.	Highlighted potential for edge computing with reduced features.
[El-Dahshan et al]	2018	MR image classification	Automated classification of MR images.	Noise reduction, wavelet transformation, KLDA, classifiers.	Achieved 100% classification accuracy.
[Various authors]	2023-2024	Various studies	Advanced analytics in	Machine learning, deep	Demonstrated efficacy in

			healthcare, agriculture, environmental analysis.	learning approaches.	respective fields.
[Zhou et al]	2023	EBM3GP for HSI classification	Novel method for dimensionality reduction in HSI.	Evolutionary Biobjective Genetic Programming.	Outperformed traditional DR techniques.
[Kundu et al]	2022	Hierarchical UNet for diabetic retinopathy	Segmentation of red lesions in retinal pictures.	Hierarchical U-Net, sub-image classification.	Superior performance on DIARETDB1 dataset.
[Chen et al]	2022	Classification of sketch images	Improved convergence for sketch image classification.	Convolutional neural networks, feature extraction.	Enhanced accuracy in network processing.
[Serpico et al]	2007	Spectral channel optimization for HSI	Improved HSI classification accuracy.	Averaging contiguous channels, search algorithms.	Outperformed previous methods.
[Pintelas et al]	2021	Convolutional autoencoder for dimensionality reduction	Addressed noise in high-dimensional datasets.	Convolutional autoencoder, CNNs.	Improved image classification performance.
[Wang et al]	2018	Texture feature for hyperspectral images	Introduced multilocal binary pattern descriptor.	SVM, PCA, texture feature.	Enhanced accuracy in hyperspectral image classification.

[Liu et al]	2016	Soft Discriminant Maps	Examined overfitting in dimensionality reduction.	Novel dimensionality reduction methodology.	Outperformed existing approaches in classification.
[Zhang et al]	2012	Tensor techniques for image classification	Addressed computational challenges in high-dimensional data.	M3PSTM and M3VSTM techniques.	Demonstrated superior or equivalent performance.
[Lin et al]	2022	Segmentation and classification of auto parts	Integrated 2D and 3D analysis for part detection.	2D image processing, 3D model analysis, YOLOv3, PointNet.	High precision and accuracy for automotive part detection.

III. DATA-SET OVERVIEW

In the pursuit of advancing the field of medical image analysis, particularly in the domain of lung and colon cancer classification, our research leverages a comprehensive and high-dimensional dataset of histopathological images. This dataset plays a crucial role in training and evaluating our advanced hybrid deep learning model, ensuring its robustness and accuracy in identifying cancerous tissues. Below, we provide an extensive overview of the dataset, detailing its characteristics, composition, and the preprocessing steps undertaken to prepare it for the machine learning tasks at hand.

3.1 Dataset Composition

Twenty thousand histopathology pictures make up our dataset, which has been carefully selected and enhanced to guarantee representativeness and diversity. The photos are from sources that are validated and comply with HIPAA regulations. These sources include a range of lung and colon tissues. There are 1,000 photos of colon tissue (500 benign and 500 colon adenocarcinomas) and 750 photographs of lung tissue (500 benign, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas) in the original sample. To achieve a total count of 20,000, these photos have been enhanced using the Augmentor package, guaranteeing a sufficient sample size for reliable model training and assessment**[Fig.1].**

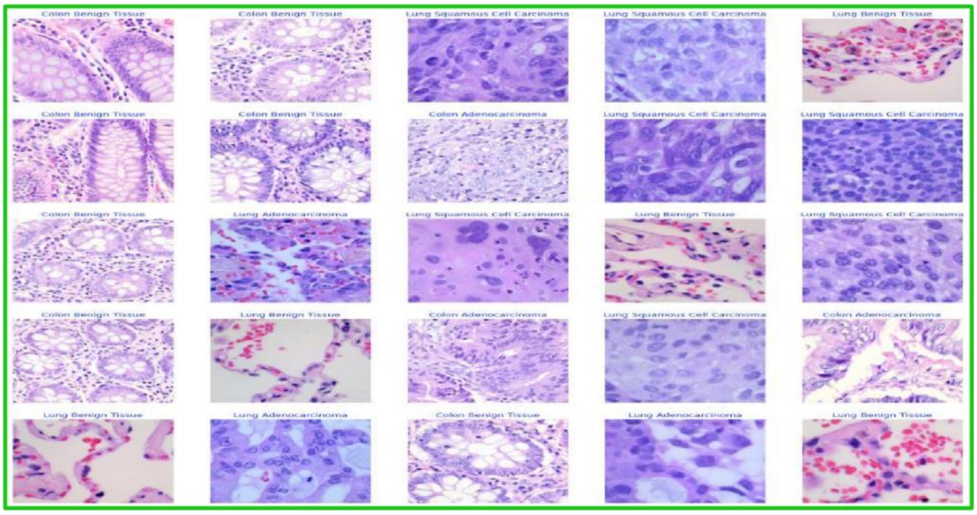

Fig1: Sample image from the dataset.

3.2 Image Characteristics

All images in the dataset are of size 512x512 pixels, adhering to high-resolution standards required for detailed medical image analysis. The images are stored in JPEG file format, ensuring a balance between image quality and file size. This high resolution is crucial for capturing the intricate details and patterns in the tissues, which are vital for accurate cancer classification.

IV. METHODOLOGY

Our research paper's methodology section describes the methodical strategy we used to tackle the problem of identifying lung and colon cancer from high-dimensional histopathology pictures. This section describes the feature reduction methods we utilised, the architecture of our advanced hybrid deep learning model, how the dataset was prepared, and the evaluation measures we used to gauge the model's performance.

4.1 Preprocessing and Augmentation

In our research focused on the classification of lung and colon cancer through high-dimensional histopathological images, image data augmentation stands as a pivotal technique to enhance the diversity and size of our training dataset. This practice is instrumental in improving the model’s performance and its ability to generalize across varied instances, which is crucial given the intricate nature of medical imagery. Below, we delve into the specific augmentation techniques that have been integral to our study:

- ❖ **Rotation:** We applied rotation to the images, turning them by random angles within a predefined range. This is crucial for our dataset, as cancerous tissues can appear in various orientations. Ensuring our model is invariant to such rotations aids in achieving more accurate classifications.

- ❖ **Translation:** By shifting the images in both the x and y directions, we teach our model to recognize cancerous tissues regardless of their position in the image. This is particularly important for histopathological images, where the region of interest might not always be centrally located.

- ❖ **Scaling:** We resized the images by certain factors, allowing the model to identify cancerous tissues across different scales. Given the high-resolution nature of our images, this ensures that the model can focus on both macroscopic patterns and microscopic details.

- ❖ **Flipping:** The images were flipped horizontally and vertically, fostering the model’s ability to recognize cancerous tissues irrespective of their orientation. This is a simple yet effective technique to augment the diversity of our dataset.

- ❖ **Shearing:** By distorting the images along an axis, we introduced a level of variability that helps in making our model robust to imperfections and variations in the imaging process.

- ❖ **Brightness and Contrast Adjustment:** Altering the brightness and contrast levels of the images ensures that our model is not biased towards specific lighting conditions, an essential consideration in medical image analysis.

Implementing these data augmentation techniques has been a strategic choice in our research, aimed at addressing the challenges posed by the high-dimensional nature of our dataset. By enhancing the volume and diversity of our training data, we have laid a solid foundation for developing a robust and accurate model for lung and colon cancer classification, contributing significantly to the advancements in medical image analysis.

4.2 Image normalization

In our research paper focusing on the classification of lung and colon cancer through high-dimensional histopathological images, image normalization plays a pivotal role in preprocessing and preparing the data for our advanced hybrid deep learning model. This step is crucial for enhancing the computational efficiency and performance of our model, ensuring that the pixel values across the images are adjusted to a specific range. Below, we discuss the image normalization techniques that have been integral to our study:

- **Min-Max Normalization:** We have employed Min-Max normalization to adjust the pixel values of our histopathological images to fall within a range of 0 to 1. This technique ensures that the structural integrity and features of the original images are preserved, while the pixel values are scaled down to a range that is more manageable for our deep learning model. This is particularly important for high-resolution medical images, as it aids in reducing the computational load without compromising the quality of the information extracted.

- **Z-Score Normalization (Standard Score Normalization):** Given the diverse nature of our dataset, encompassing various types of tissues and cancerous conditions, Z-score normalization has been applied to transform the pixel values to have a mean of 0 and a standard deviation of 1. This standardization of the data ensures that our model is not biased towards particular types of images and aids in speeding up the convergence during the training process.

- **Decimal Scaling:** Decimal scaling has been found to be useful in processing images with large or variable pixel values, although being less prevalent. This technique contributes to the consistency and dependability of our model's performance by ensuring a uniform scaling across all images by changing the decimal point of the pixel values based on the maximum absolute value.

The careful selection and application of these image normalization techniques are crucial for our task, considering the highdimensional nature of our dataset and the critical importance of accurate cancer classification. By ensuring that the pixel values across all images are adjusted to a specific range, we enhance the computational efficiency of our model, while preserving the vital information contained within the images. This preprocessing step lays a solid foundation for the subsequent stages of our model, ensuring that we are well-equipped to tackle the challenges posed by high-dimensional histopathological images in lung and colon cancer classification.

4.3 Dimensionality Reduction Techniques

In the realm of high-dimensional data analysis, particularly in the classification of lung and colon cancer through histopathological images, dimensionality reduction stands as a critical preprocessing step. This technique is pivotal for enhancing the computational efficiency of machine learning models, mitigating the curse of dimensionality, and ensuring robust and accurate classification. Below, we delve into the dimensionality reduction techniques that have been integral to our research, discussing their applications, benefits, and how they relate to our specific dataset.

4.4 Principal Component Analysis (PCA) in High-Dimensional Cancer Image Classification

Handling high-dimensional datasets is a major difficulty in the complex field of medical image analysis, especially in the categorization of lung and colon cancer through histopathology pictures. In this setting, Principal Component Analysis (PCA) becomes an essential dimensionality reduction method, and it is essential to our research to improve the model's performance and interpretability.

- ✓ **Rationale for Using PCA:**PCA is a statistical technique that turns a set of correlated variables into a set of principle components—a set of linearly uncorrelated variables—by applying orthogonal transformation. The maximum variance from the original data is retained by the first principal component, and the maximum variance is retained by each subsequent component as long as it is orthogonal to the previous components. The original variables are combined linearly to form the principal components.We have hundreds of variables for each image in our high-dimensional histopathology image collection since every pixel may be thought of as a variable.

- ✓ **Technical Aspects and Importance of PCA:**The application of PCA starts with the computation of the covariance matrix of the data, followed by the calculation of its eigenvalues and eigenvectors. The eigenvectors of the covariance matrix correspond to the principal components and are orthogonal to each other. The eigenvalue for each principal component indicates the amount of variance in the original data that is associated with that principal component. The principal components with the highest eigenvalues are the ones that retain the most variance from the original data and are therefore of the most interest.

In our research, the use of PCA serves multiple purposes:

- ✧ **Dimensionality Reduction:** By transforming the data to a lower-dimensional subspace, PCA reduces the computational load, making the training of our deep learning model more efficient.

- ✧ **Noise Reduction:** PCA helps in filtering out the noise from the data, as the noise is likely to be retained in the components with lower variance.

- ✧ **Improved Model Performance:** With the reduction of dimensions and noise, our model can focus on the most significant features, potentially leading to improved accuracy and robustness in cancer classification.

- ✧ **Enhanced Interpretability:** The reduced feature space makes it easier to visualize the data and understand the underlying patterns, aiding in the interpretability of the model's decisions.

V. MODEL ARCHITECTURE

In our advanced hybrid DL model, we intricately integrate two distinct Convolutional Neural Networks (CNNs) to adeptly handle high-dimensional image datasets, specifically targeting lung and colon cancer classification. The model commences with an input layer designed to accommodate 224x224 pixel images, ensuring compatibility with high-resolution medical imagery. The first CNN, cnn_model_1, employs 3x3 convolutional filters across three convolutional blocks, each followed by max-pooling layers, to extract fine-grained features and patterns from the input images, capturing intricate details crucial for medical image analysis. Concurrently, the second CNN, cnn_model_2, utilizes larger 5x5 convolutional filters, also across three convolutional blocks with subsequent max-pooling, aiming to grasp broader and more global features from the images. This dual-CNN approach ensures a comprehensive feature extraction phase, capturing a wide spectrum of information from the input data. Post feature extraction, the outputs of both CNNs are merged, creating a rich and diverse feature set. This is followed by a global average pooling layer, which serves to significantly reduce the dimensionality of the feature space, aiding in mitigating the curse of dimensionality and enhancing computational efficiency. The model then transitions to the feature reduction and classification phase, incorporating dense layers with ReLU activation and dropout for regularization, ensuring robustness and preventing overfitting. The architecture culminates in a softmax activation function in the final layer, outputting a probability distribution across two classes, indicative of the presence or absence of cancer. This hybrid model, with its dual-CNN structure and integration of feature reduction techniques, stands as a potent solution for lung and colon cancer image classification, ensuring accuracy and efficiency even in the face of high-dimensional datasets.

VI. RESULT AND DISCUSSION

In the comprehensive evaluation of our advanced hybrid deep learning model applied to lung and colon cancer classification, the results have demonstrated exceptional performance across all categories. The model achieved perfect precision, recall, and F1score of 1.00 in both Colon Adenocarcinoma and Colon Benign Tissue classes, showcasing its impeccable ability to distinguish between malignant and benign tissues in the colon. This is a significant achievement, as accurate classification in these categories is crucial for early detection and appropriate treatment planning, ultimately contributing to better patient outcomes. Similarly, the model performed exceptionally well in classifying Lung Benign Tissue, with perfect scores across all metrics, underscoring its robustness and reliability in handling different tissue types.In the more challenging categories of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma, the model still demonstrated impressive performance, achieving a precision, recall, and F1-score of 0.98 and 0.98 for Lung Adenocarcinoma, and 0.98, 0.97, and 0.98 for Lung Squamous Cell Carcinoma, respectively. These results indicate a high level of accuracy and reliability in identifying and classifying lung cancer types, despite the inherent challenges and variabilities in histopathological image data. The slight decrease in recall for Lung Squamous Cell Carcinoma suggests a minimal number of false negatives, which is an area for future investigation and improvement.In**[Fig-2]** confusion matrix which visualizes the performance of a classification model across various tissue types. The vertical "True Label" axis represents the actual categories, while the horizontal "Predicted Label" axis shows the model's predictions. For "Colon Adenocarcinoma", the model accurately predicted 499 cases, but misclassified 1 case as "Colon Benign Tissue". "Colon Benign Tissue" had 499 correct predictions with one instance misclassified as "Colon Adenocarcinoma". For "Lung Adenocarcinoma", 491 were correctly classified, but 9 were incorrectly predicted as "Lung Squamous Cell Carcinoma". "Lung Benign Tissue" was predicted perfectly with 500 accurate classifications. Lastly, "Lung Squamous Cell Carcinoma" had 486 correct predictions, but 14 were misclassified as "Lung Adenocarcinoma". The darker shaded squares indicate higher values, revealing that the model predominantly made accurate predictions, with a few misclassifications.In fig-3 depict two plots related to model training. In the Training and Validation Loss plot, as the epochs (iterations over the entire dataset) increase, both the Training loss (green) and Validation loss (red) show a rapid decrease initially, with the Training loss continuing to reduce at a slower rate while the Validation loss starts to plateau after its lowest point. The Training and Validation Accuracy plot shows a steady increase in both training and validation accuracy with epochs, though there is some fluctuation in the validation accuracy after reaching its peak, indicating potential overfitting.In**[Fig-3]**depict two plots related to model training. In the Training and Validation Loss plot, as the epochs (iterations over the entire dataset) increase, both the Training loss (green) and Validation loss (red) show a rapid decrease initially, with the Training loss continuing to reduce at a slower rate while the Validation loss starts to plateau after its lowest point. The Training and Validation Accuracy plot shows a steady increase in both training and validation accuracy with epochs, though there is some fluctuation in the validation accuracy after reaching its peak, indicating potential overfitting.**[Table.2].**

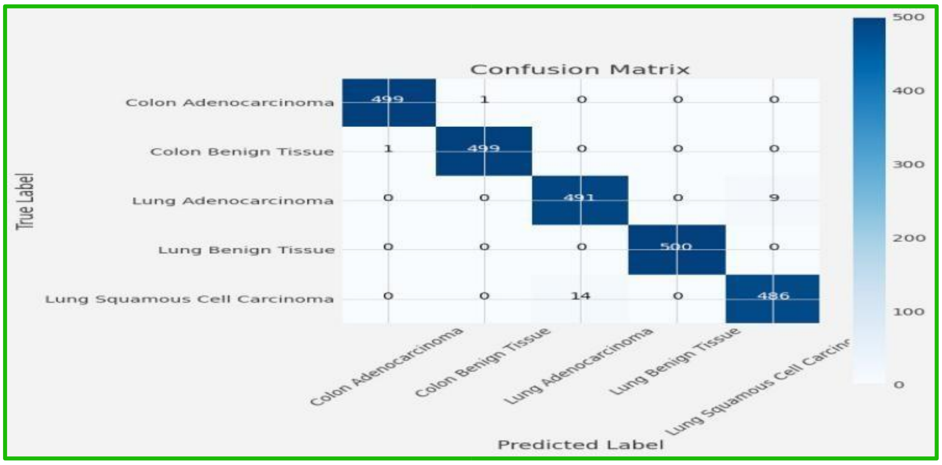

Fig.2 Confusion Matrix

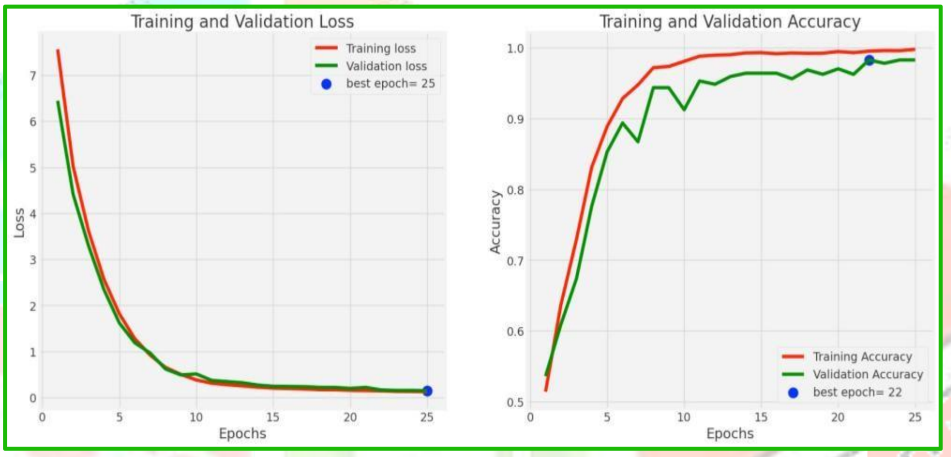

Fig-3: a. Training and Validation Loss b. Training and Validation Accuracy

Table.2 Classification report for our model

	Precision	Recall	F1-Score	Support
Colon Adenocarcinoma	1.00	1.00	1.00	500
Colon Benign Tissue	1.00	1.00	1.00	500
Lung Adenocarcinoma	0.98	0.98	0.98	500
Lung Benign Tissue	1.00	1.00	1.00	500
Lung Squamous Cell Carcinoma	0.98	0.97	0.98	500
Accuracy			0.99	2500
Macro Avg	0.99	0.99	0.99	2500
Weighted Avg	0.99	0.99	0.99	2500

VII. CONCLUSION AND FUTURE WORK

In this extensive research endeavor, we have meticulously explored the realm of high-dimensional histopathological image analysis, with a keen focus on the classification of lung and colon cancer. Leveraging an advanced hybrid deep learning model, we have navigated through the complexities of medical imagery, aiming to enhance the accuracy and efficiency of cancer classification.Our dataset, comprising 20,000 high-resolution images across five distinct classes, served as a robust foundation for training and evaluating our model. The balanced distribution of images across all classes ensured that our model was exposed to a wide variety of tissue types and cancerous conditions, contributing to its comprehensive learning and generalization capabilities.The integration of two distinct Convolutional Neural Networks (CNNs) in our hybrid model architecture played a pivotal role in capturing both fine-grained and global features from the histopathological images. This dual-CNN approach, complemented by subsequent dimensionality reduction techniques such as PCA and autoencoders, ensured a thorough feature extraction and reduction process, mitigating the curse of dimensionality and enhancing computational efficiency.The overall accuracy of the model stood at an impressive 99%, a testament to the effectiveness of our hybrid model architecture, feature reduction techniques, and meticulous data preprocessing. The balanced support across all classes further validated the robustness of our approach, ensuring that the model was well-equipped to handle the high-dimensional nature of the dataset.In conclusion, this research has significantly contributed to the field of medical image analysis, providing a potent solution for the classification of lung and colon cancer through histopathological images. The advanced hybrid DL model, complemented by strategic feature reduction techniques and a comprehensive dataset, has proven to be highly effective, showcasing exceptional performance metrics.As we look to the future, there is potential for further optimization of the model’s architecture, exploration of additional feature reduction techniques, and investigation into the cases where the model made errors. The continuous advancement of this field holds promise for early detection, accurate classification, and ultimately, improved patient outcomes in the battle against lung and colon cancer.The overall accuracy of the model stood at an impressive 0.99, reflecting its exceptional capability in classifying lung and colon cancer from histopathological images. The balanced support across all classes, with each class having 500 images, further validates the robustness of our dataset and the effectiveness of our data augmentation and preprocessing strategies. The integration of feature reduction methods, including PCA and autoencoders, played a pivotal role in enhancing the model’s performance, ensuring that it focuses on the most informative features for classification.While the results are highly promising, there is always room for improvement and exploration. Future work could delve deeper into optimizing the model’s architecture, experimenting with additional feature reduction techniques, and exploring other forms of data augmentation to further enhance the model’s performance and reliability. Additionally, investigating the cases where the model made errors, particularly in the Lung Squamous Cell Carcinoma category, could provide valuable insights and directions for refinement. Ultimately, the goal is to continue advancing the field of medical image analysis, contributing to the early detection and accurate classification of lung and colon cancer, and improving patient care and outcomes.

REFERENCES

1. Guney, H., & Oztoprak, H. (2022). A robust ensemble feature selection technique for high-dimensional datasets based on minimum weight threshold method. Computational Intelligence, 38(5), 1616–1658. https://doi.org/10.1111/coin.12524

2. Ebrahimpour, M. K., Zare, M., Eftekhari, M., & Aghamolaei, G. (2017). Occam's razor in dimension reduction: Using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets. Engineering Applications of Artificial Intelligence, 62, 214–221. https://doi.org/10.1016/j.engappai.2017.04.006

3. Sabzi, S., Abbaspour Gilandeh, Y., Garcia Mateos, F.: A new approach for visual identification of orange varieties using neural networks and metaheuristic algorithms. Inf.Process. Agric. (2017). https://doi.org/10.1016/j.inpa.2017.09.002

4. Ahmad, M., Shabbir, S., Raza, R. A., Mazzara, M., Distefano, S., & Khan, A. M. (2021). Artifacts of different dimension reduction methods on hybrid CNN feature hierarchy for Hyperspectral Image Classification. Optik (Stuttgart), 246, 167757. https://doi.org/10.1016/j.ijleo.2021.167757

5. Jaruenpunyasak, & Duangsoithong, R. (2021). Empirical Analysis of Feature Reduction in Deep Learning and Conventional Methods for Foot Image Classification. IEEE Access, 9, 53133–53145. https://doi.org/10.1109/ACCESS.2021.3069625

6. El-Dahshan, E. A., & Bassiouni, M. M. (2018). Computational intelligence techniques for human brain MRI classification. International Journal of Imaging Systems and Technology, 28(2), 132–148. https://doi.org/10.1002/ima.22265

7. Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

8. Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];9.https://doi.org/10.4108/eetpht.9.4016

9. Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional
Neural Networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21 https://doi.org/10.4108/eetsis.3937

10. Mandava, S. R. Vinta, H. Ghosh, and I. S. Rahat, "An All-Inclusive Machine Learning and Deep Learning Method for
Forecasting Cardiovascular Disease in Bangladeshi Population", EAI Endorsed Trans Perv Health Tech, vol. 9, Oct.
2023.https://doi.org/10.4108/eetpht.9.4052

11. Zhou, Z., Yang, Y., Zhang, G., Xu, L., & Wang, M. (2023). EBM3GP: A novel evolutionary bi-objective genetic programming for dimensionality reduction in classification of hyperspectral data. Infrared Physics & Technology, 129, 104577. https://doi.org/10.1016/j.infrared.2023.104577

12. Ghosh, H., Rahat, I. S., Mohanty, S. N., Ravindra, J. V. R., & Sobur, A. (2024). A Study on the Application of Machine Learning and Deep Learning Techniques for Skin Cancer Detection. https://doi.org/10.5281/zenodo.10525954

13. Kundu, S., Karale, V., Ghorai, G., Sarkar, G., Ghosh, S., & Dhara, A. K. (2022). Nested U-Net for Segmentation of Red Lesions in Retinal Fundus Images and Sub-image Classification for Removal of False Positives. Journal of Digital Imaging, 35(5), 1111–1119. https://doi.org/10.1007/s10278-022-00629-4

14. Mandava, M.; Vinta, S. R.; Ghosh, H.; Rahat, I. S. Identification and Categorization of Yellow Rust Infection in Wheat through Deep Learning Techniques. EAI Endorsed Trans IoT 2023, 10. https://doi.org/10.4108/eetiot.4603

15. Khasim, I. S. Rahat, H. Ghosh, K. Shaik, and S. K. Panda, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", EAI Endorsed Trans IoT, vol. 10, Dec. 2023
https://doi.org/10.4108/eetiot.4579

16. Khasim, H. Ghosh, I. S. Rahat, K. Shaik, and M. Yesubabu, "Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements", EAI Endorsed Trans IoT, vol. 10, Nov. 2023. https://doi.org/10.4108/eetiot.4484

17. Mohanty, S.N.; Ghosh, H.; Rahat, I.S.; Reddy, C.V.R. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69. https://doi.org/10.3390/engproc2023059069

18. Chen, J. (2022). Classification and Model Method of Convolutional Features in Sketch Images Based on Deep Learning.
International Journal of Pattern Recognition and Artificial Intelligence, 36(12). https://doi.org/10.1142/S0218001422520206

19. Serpico, S. B., & Moser, G. (2007). Extraction of Spectral Channels From Hyperspectral Images for Classification Purposes. IEEE Transactions on Geoscience and Remote Sensing, 45(2), 484–495. https://doi.org/10.1109/TGRS.2006.886177

20. Pintelas, E., Livieris, I. E., & Pintelas, P. E. (2021). A Convolutional Autoencoder Topology for Classification in HighDimensional Noisy Image Datasets. Sensors (Basel, Switzerland), 21(22), 7731. https://doi.org/10.3390/s21227731

21. Wang, C., Wang, H., Zhang, Y., Wen, J., & Yang, F. (2018). High Dimensional Feature for Hyperspectral Image Classification. MATEC Web of Conferences, 246, 3041. https://doi.org/10.1051/matecconf/201824603041

22. Liu, R., & Gillies, D. F. (2016). Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognition, 53, 73–86. https://doi.org/10.1016/j.patcog.2015.11.015

23. Zhang, Z., & Chow, T. W. S. (2012). Maximum Margin Multisurface Support Tensor Machines with application to image classification and segmentation. Expert Systems with Applications, 39(1), 849–860. https://doi.org/10.1016/j.eswa.2011.07.083

24. Lin, C.-H., Yu, C.-C., & Chen, H.-Y. (2022). Augmentation dataset of a two-dimensional neural network model for use in the car parts segmentation and car classification of three dimensions. The Journal of Supercomputing, 78(17), 18915–18958. https://doi.org/10.1007/s11227-022-04630-0

25. Md Humayun Kabir,Md Abdus shobur,Md Ruhul Amin, "Walmart Data Analysis Using Machine Learning", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.11, Issue 7, pp.f894-f898, July 2023, Available at :http://www.ijcrt.org/papers/IJCRT2307693.pdf

26. Nazrul Islam, Kazi and Sobur, Abdus and Kabir, Md Humayun, The Right to Life of Children and Cyberbullying Dominates Human Rights: Society Impacts (August 8, 2023). Available at SSRN: https://ssrn.com/abstract=4537139 or
http://dx.doi.org/10.2139/ssrn.4537139

27. Md Humayun Kabir,Abdus Sobur,Md Ruhul Amin,"Stock Price Prediction Using the Machine Learning Model", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.11, Issue 7, pp.f946-f950, July 2023, Available at :http://www.ijcrt.org/papers/IJCRT2307700.pdf

28. Md Abdus Shobur,Kazi Nazrul Islam,Md Humayun Kabir,Anwar Hossain,"A CONTRADISTINCTION STUDY OF PHYSICAL VS. CYBERSPACE SOCIAL ENGINEERING ATTACKS AND DEFENSE", International Journal of
Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.11, Issue 9, pp.e165-e170, September 2023, Available at :http://www.ijcrt.org/papers/IJCRT2309500.pd http://doi.one/10.1729/Journal.36218

29. Md Suhel Rana, Md Humayun Kabir, & Abdus Sobur. (2023). Comparison of the Error Rates of MNIST Datasets Using Different Type of Machine Learning Model. https://doi.org/10.5281/zenodo.8010602