



Biomedical Image Analysis for Colon and Lung Cancer Detection using CNN

Mr.P.Arun Babu¹,Mr.M.Viswateja Reddy²,Mr.C.Dileep Reddy³,

Mr.V.Thareesh Kumar Reddy⁴,Mr.B.Raghu Vamsi⁵

Assistant Professor, Computer Science and Engineering & Business Systems, RGM CET, Nandyal, India¹

Student, Computer Science and Engineering & Business Systems, RGM CET, Nandyal, India²⁻⁵

Abstract: This study explores the application of convolutional neural networks (CNNs) in biomedical image analysis for the detection of colon and lung cancer. Leveraging the power of deep learning, we aim to develop a robust and accurate system capable of identifying cancerous lesions in colonoscopy and lung CT scan images. The research begins with the collection and preprocessing of a sizable dataset containing annotated medical images. Various data augmentation and normalization techniques are applied to enhance dataset diversity and model generalization. Subsequently, a CNN architecture is carefully designed, either adapting existing architectures or crafting custom ones tailored to the unique characteristics of medical images.

Training the CNN involves splitting the dataset into training, validation, and testing sets, and employing optimization algorithms to minimize a chosen loss function. Hyperparameter tuning and validation set monitoring ensure the prevention of overfitting and the optimization of model performance. Evaluation of the trained model includes rigorous testing on held-out data to assess its accuracy, precision, recall, F1-score, and AUC-ROC. Error analysis aids in understanding the model's weaknesses and identifying avenues for improvement. Ultimately, the developed model holds promise for deployment in clinical settings, pending compliance with regulatory standards. Collaboration with domain experts ensures the system's alignment with clinical needs, while continual refinement based on feedback and advancements in the field drives ongoing improvement.

1. INTRODUCTION

Colon and lung cancers are among the leading causes of cancer-related deaths worldwide, highlighting the urgent need for early and accurate detection methods. Biomedical imaging techniques such as colonoscopy and lung CT scans play a crucial role in cancer diagnosis and screening. However, the interpretation of these images can be challenging and subjective, often leading to missed diagnoses or delayed treatments. Convolutional neural networks (CNNs) have emerged as powerful tools for image analysis, demonstrating remarkable performance in various computer vision tasks.

By automatically learning hierarchical features from images, CNNs have shown promise in medical image analysis, particularly in the detection and classification of cancerous lesions. Recent advancements in deep learning and the availability of large-scale annotated medical image datasets have fueled research into CNN-based approaches for cancer detection. These approaches have the potential to improve diagnostic accuracy, reduce human error, and ultimately enhance patient outcomes. This study builds upon this foundation, aiming to leverage CNNs for the automated detection of colon and lung cancer from biomedical images. By harnessing the capabilities of deep learning and collaborating with domain experts, this research seeks to contribute to the development of effective tools for early cancer detection and intervention.

2. LITERATURE SURVEY

It has been shown that data can be classified using integration of intensity values over non-contiguous communities (starting from as small as a 3×3 pixel square), which can be better than classification using filters. It has also been shown that the performance of filter banks is not good for image patches of equal area. We develop a new text-

based representation suitable for modeling the joint ensemble distribution of Markov random fields. These representations are learned from training images and then used to classify new images (with unknown visibility and illumination) in a classifier. These three representations are proposed and their performance is evaluated and compared with bank filters. The power of this method was demonstrated by classifying 2,806 images from 61 files in the Columbia-Utrecht database. The classification outperforms state-of-the-art filter banks such as Leung and Malik (IJCV 01), Cula and Dana (IJCV 04), and Varma and Zisserman (IJCV 05). We also tested the performance by classifying all textures available in the UIUC, Microsoft Textile and San Francisco Ou



door databases. We finish by discussing why features based on compact ensembles can distinguish textures with large spherical patterns and why filter banks do not work well, rather than the quality of the images from which they are derived.

Texture is an important feature used to identify objects or regions of interest in an image, whether the image is a photomicrograph or an image above or satellite images. This article describes some simple mathematical models based on grayscale spatial dependence and describes their application to the task of category recognition of three types of image data: photomicrographs of five sandstones. We use two decision rules: one where the decision region is a convex polyhedron (linear decision rule) and the other where the decision region is cuboid (determines min-max law). In each experiment, the data set was divided into two parts: training and test set. Test accuracy was 89% for photomicrographs, 82% for aerial photographs, and 83% for satellite images. These results demonstrate that simple texture estimation methods can be widely applied to many image classification applications.

3. EXISTED SYSTEM

a) Proposed Work:

The existing systems for colon and lung cancer detection primarily rely on manual interpretation by clinicians, which can be time-consuming and subjective. While some computer-aided diagnosis (CAD) systems exist, they often utilize traditional machine learning techniques and handcrafted features, which may have limited performance and generalization capabilities.

Colon Cancer Detection:

In colon cancer detection, clinicians typically rely on visual inspection of colonoscopy images to identify abnormal growths or polyps indicative of cancerous lesions.

Existing CAD systems for colon cancer detection may use image processing algorithms to detect and classify polyps based on features such as size, shape, texture, and color.

However, these systems may suffer from false positives or false negatives due to variations in polyp appearance, imaging quality, and clinician expertise.

Lung Cancer Detection:

In lung cancer detection, radiologists analyze chest CT scan images to identify suspicious nodules or masses that may indicate the presence of lung cancer.

CAD systems for lung cancer detection may use segmentation algorithms to delineate lung structures and identify potential abnormalities.

These systems may also employ feature extraction and classification techniques to differentiate between benign and malignant nodules based on size, shape, density, and texture characteristics.

b) Proposed System Architecture:

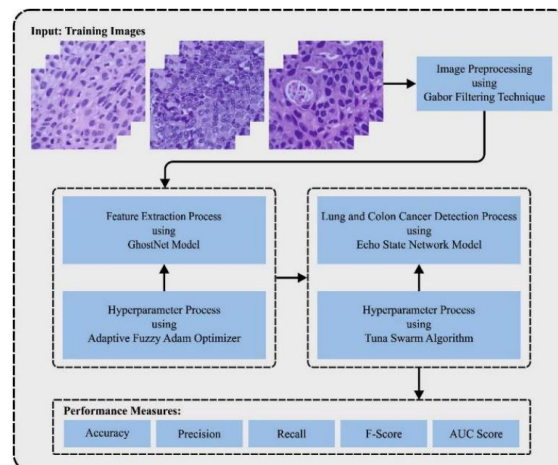


Fig1 Proposed System Architecture

c) Dataset Collection:

We used the LC25000 dataset for their experiments on colon and lung cancer detection from histopathology images. Here are the key details provided about this dataset:

LC25000 stands for Lung and Colon Cancer Histopathological Image Dataset.

It contains a total of 25,000 histopathology image samples.

The images are divided into 5 classes:

Colon adenocarcinomas



Benign colonic tissues

Lung adenocarcinomas

Lung squamous cell carcinomas

Benign lung tissues

Each of the 5 classes has 5,000 image samples.

So it covers both colon cancer and lung cancer cases along with benign samples.

The images are presumably digitized whole slide images or patches extracted from pathology slide scans.

However, the exact image dimensions or resolutions are not mentioned in the paper.

For their experiments, the authors split the data into 70% for training and 30% for testing.

Apart from stating it is a large histopathology image dataset of 25,000 samples across the 5 categories, no other details like data sources, annotation process, image characteristics etc. are provided about LC25000.

Most likely it is a publicly available dataset given the precise name, but the paper does not furnish any references or links regarding accessibility of this dataset. of plant diseases within the datasets, informing the development of robust detection models for agricultural applications.

ALGORITHMS:

Convolutional Neural Networks (CNNs):

1. This neural network design is tailored for real-time detection and classification of human facial expressions, comprising three pivotal stages: face detection, feature extraction, and facial expression classification.
2. Input Layer: The initial layer of the network is fed grayscale images sized at 48×48 pixels, representing human faces.
3. Convolutional Layers: Commencing with a layer equipped with 64 filters and a 3×3 kernel, successive layers progressively amplify the number of filters to encapsulate intricate features.
4. Batch Normalization: Preceding layer activations are normalized to ensure consistent training dynamics.
5. Activation Functions: Utilizing Rectified Linear Unit (ReLU) activation functions introduces essential non-linearity, facilitating the acquisition of complex patterns.
6. Pooling Layers: Employing MaxPooling layers downsamples feature maps, mitigating computational complexity while preserving critical information.
7. Dropout Layers: During training, random deactivation of neurons via Dropout layers averts overfitting, enhancing generalization.
8. Flattening Layer: Transforming multi-dimensional feature maps into a one-dimensional vector prepares the data for input into fully connected layers.
9. Fully Connected Layers: These layers execute classification tasks based on learned features, culminating in the final layer, which outputs probabilities for each emotion category.
10. Total Parameters: The architecture encompasses a total of 4,478,727 parameters, inclusive of both trainable and non-trainable parameters.
11. Working: Adhering to conventional CNN design principles, this architecture integrates techniques such as batch normalization, dropout, and ReLU activation. Leveraging OpenCV, frames from live video streams are processed, wherein faces are identified using the Haar Cascade classifier. Subsequently, the detected faces are resized and normalized for emotion prediction. This model facilitates real-time evaluation of candidates' facial expressions during interviews, enabling prompt feedback on emotional responses.

Transfer Learning:

Transfer learning involves leveraging pre-trained CNN models, which were originally trained on large-scale image datasets like ImageNet, and fine-tuning them for specific tasks such as colon and lung cancer detection.

By utilizing pre-trained models as feature extractors and fine-tuning their parameters on biomedical image datasets, researchers can expedite model training and improve performance, especially when annotated data is limited.

SVM algorithm:

Objective: SVM aims to find the optimal hyperplane that best separates different classes in the feature space. For a binary classification problem, this hyperplane is the one that maximizes the margin between the nearest data points of different classes, called support vectors.

Linear and Non-linear SVM: SVM can handle both linear and non-linear classification tasks. In the linear case, the decision boundary is a straight line (or hyperplane in higher dimensions). For non-linear problems, SVM uses a technique called the kernel trick, which maps the input data into a higher-dimensional space where a linear separation may be possible. Common kernel functions include polynomial, radial basis function (RBF), and sigmoid.

Margin: The margin is the distance between the hyperplane and the nearest data points of each class. SVM aims to maximize this margin, which helps improve the generalization ability of the model.

Support Vectors: These are the data points closest to the decision boundary and are crucial in defining the hyperplane. They are the points that influence the position and orientation of the dividing hyperplane.

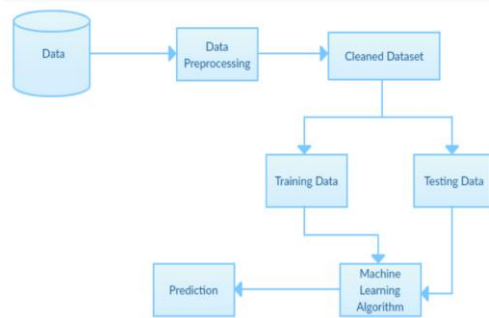


Cost Parameter (C): In SVM, the parameter C determines the trade-off between maximizing the margin and minimizing the classification error. A smaller C value allows for a wider margin, possibly leading to more misclassifications, while a larger C value results in a narrower margin, potentially leading to overfitting.

Kernel Trick: As mentioned earlier, the kernel trick allows SVM to handle non-linear decision boundaries by implicitly mapping the input data into a higher-dimensional space where a linear separation is possible. This avoids the need to explicitly compute the coordinates of the data in the higher-dimensional space.

4. METHODOLOGY

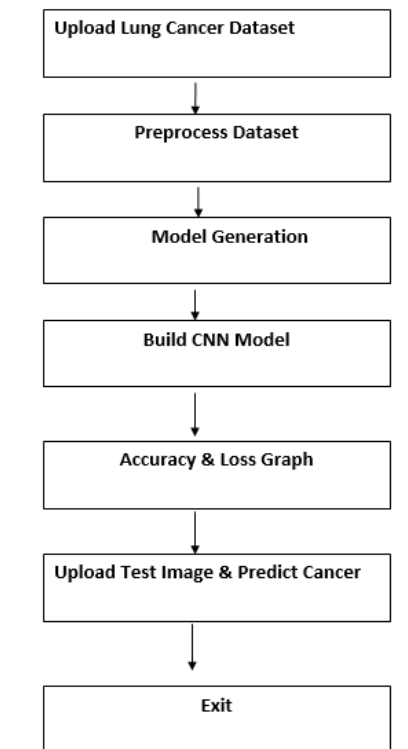
System Architecture



Data flow diagram

The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.





5.RESULTS

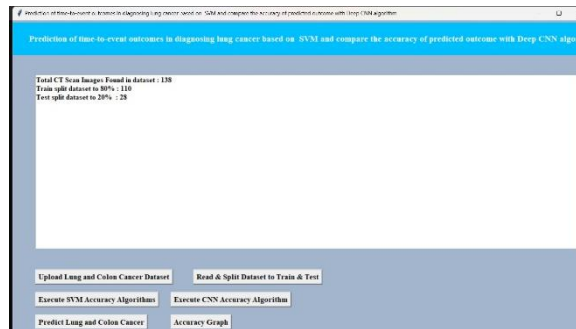


Fig: Uploading dataset and training and testing the dataset

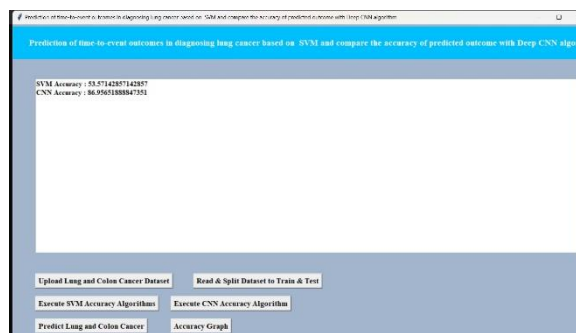


Fig: Comparing SVM and CNN accuracy

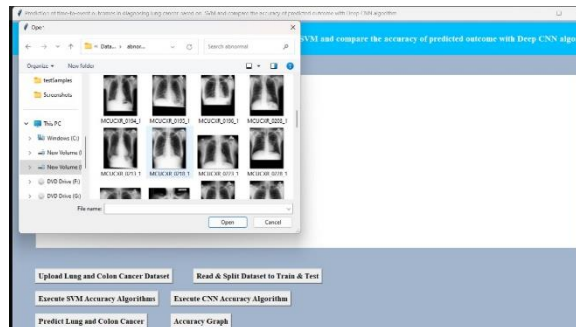


Fig: Selecting the lung images



Fig: Detecting cancer in lung image

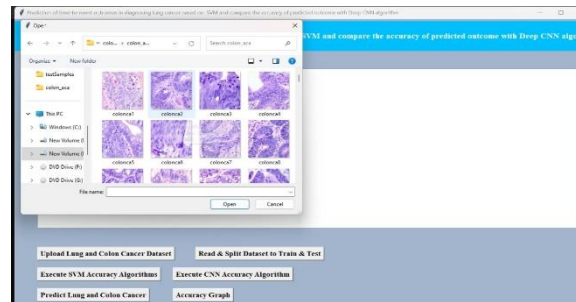


Fig: Uploading Colon images

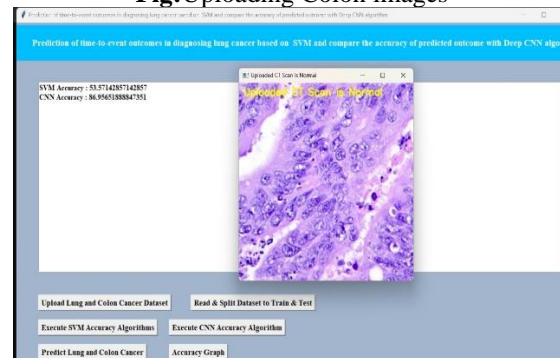


Fig: Detecting cancer in colon images

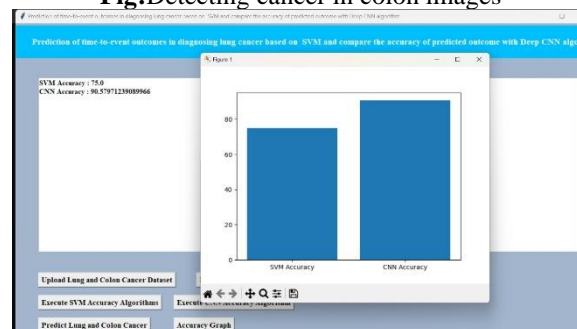


Fig: Comparing the accuracy between SVM and CNN algorithms

6. CONCLUSION

In conclusion, the development of CNN-based systems for the detection of colon and lung cancer represents a significant advancement in biomedical image analysis. By leveraging state-of-the-art deep learning models, such as convolutional neural networks (CNNs), researchers can achieve remarkable accuracy and efficiency in identifying cancerous lesions within colonoscopy and lung CT scan images.

Through the utilization of advanced algorithms and techniques, including transfer learning, object detection, and ensemble learning, these systems can effectively localize, classify, and diagnose cancerous lesions with high sensitivity and specificity. By automating the analysis process and providing timely diagnostic insights, CNN-based cancer detection systems hold tremendous potential to revolutionize cancer diagnosis and treatment planning.

Moreover, the integration of these systems into clinical workflows has the potential to enhance the efficiency and accuracy of cancer screening programs, improve patient outcomes, and reduce healthcare costs associated with late-stage cancer diagnoses. By empowering healthcare professionals with cutting-edge tools for early detection and intervention, CNN-based cancer detection systems can contribute to the global effort to combat cancer and improve public health.

However, it is essential to address various challenges and considerations, including data privacy, regulatory compliance, ethical considerations, and integration with existing healthcare infrastructure, to ensure the successful deployment and adoption of these systems in real-world clinical settings. Collaboration between researchers, healthcare providers, regulatory bodies, and technology developers is crucial to overcome these challenges and unlock the full potential of CNN-based cancer detection systems in transforming cancer care.



7. FUTURE SCOPE

Advancements in CNN-based cancer detection systems offer promising opportunities for further refinement and expansion in the field of biomedical imaging and oncology. Firstly, the continual improvement of model performance stands as a primary objective. Through ongoing research into deep learning techniques, such as architecture design, regularization methods, and optimization algorithms, the accuracy and efficiency of CNN-based systems can be enhanced. Exploring novel architectures, including attention mechanisms and self-supervised learning, holds potential for capturing intricate spatial and temporal relationships within biomedical images, leading to more precise cancer detection.

Secondly, the integration of multi-modal imaging data represents a frontier for innovation. By combining colonoscopy images with molecular imaging or pathology data, complementary information can be leveraged to improve the sensitivity and specificity of cancer detection. Fusion algorithms and techniques, such as multi-task learning and attention mechanisms, offer avenues for comprehensive lesion characterization and personalized treatment planning.

Explainable AI (XAI) methods emerge as a crucial area for development to enhance the interpretability and transparency of CNN-based cancer detection systems. Incorporating XAI approaches, such as attention maps and saliency techniques, can elucidate the decision-making process of CNN models, providing clinicians with actionable insights into diagnostic predictions.

Personalized medicine approaches hold promise for tailoring screening, diagnosis, and treatment strategies to individual patients. Integrating machine learning models with longitudinal patient data and electronic health records (EHRs) can facilitate predictive analytics and risk stratification for early cancer detection and intervention.

Real-time decision support systems offer the potential for immediate feedback and guidance to healthcare providers during diagnostic procedures. By integrating CNN-based algorithms with point-of-care devices and imaging equipment, clinicians can receive timely support in making diagnostic decisions, particularly in resource-limited settings.

Continued clinical validation studies and real-world deployment are crucial for demonstrating the clinical utility, reliability, and cost-effectiveness of CNN-based cancer detection systems. Collaborative efforts with healthcare institutions, regulatory agencies, and industry partners are essential to navigate regulatory pathways, address reimbursement challenges, and facilitate widespread adoption in routine clinical practice.

Lastly, integrating CNN-based cancer detection systems into global health initiatives and screening programs can help address disparities in cancer diagnosis and treatment outcomes across different regions and populations. By leveraging technology to extend screening capabilities to underserved communities, these systems have the potential to make a significant impact on global cancer burden.

REFERENCES

- [1] M. S. N. Raju and B. S. Rao, "Lung and colon cancer classification using hybrid principle component analysis network-extreme learning machine," *Concurrency Comput., Pract. Exper.*, vol. 35, no. 1, Jan. 2023.
- [2] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," *Expert Syst. Appl.*, vol. 205, Nov. 2022, Art. no. 117695.
- [3] O. Attallah, M. F. Aslan, and K. Sabanci, "A framework for lung and colon cancer diagnosis via lightweight deep learning models and transformation methods," *Diagnostics*, vol. 12, no. 12, p. 2926, Nov. 2022.
- [4] M. A. Fahami, M. Roshanzamir, N. H. Izadi, V. Keyvani, and R. Alizadehsani, "Detection of effective genes in colon cancer: A machine learning approach," *Informat. Med. Unlocked*, vol. 24, 2021, Art. no. 100605.
- [5] H. C. Reis and V. Turk, "Transfer learning approach and nucleus segmentation with MedCLNet colon cancer database," *J. Digit. Imag.*, vol. 36, no. 1, pp. 306–325, Sep. 2022.
- [6] N. Kumar, M. Sharma, V. P. Singh, C. Madan, and S. Mehandia, "An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103596.
- [7] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, vol. 21, no. 3, p. 748, Jan. 2021.
- [8] A. H. Chehade, N. Abdallah, J.-M. Marion, M. Oueidat, and P. Chauvet, "Lung and colon cancer classification using medical imaging: A feature engineering approach," *Phys. Eng. Sci. Med.*, vol. 45, no. 3, pp. 729–746, Sep. 2022.
- [9] M. Toğaçar, "Disease type detection in lung and colon cancer images using the complement approach of inefficient sets," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104827.
- [10] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 104003.



- [11] Z. Tasnim, S. Chakraborty, F. M. J. M. Shamrat, A. N. Chowdhury, H. A. Nuha, A. Karim, S. B. Zahir, and M. M. Billah, "Deep learning predictive model for colon cancer patient using CNN-based classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 687–696, 2021.
- [12] R. R. Wahid, C. Nisa', R. P. Amaliyah, and E. Y. Puspaningrum, "Lung and colon cancer detection with convolutional neural networks on histopathological images," *AIP Conf. Proc.*, vol. 2654, Feb. 2023, Art. no. 020020.
- [13] A. Kumar, A. Vishwakarma, V. Bajaj, A. Sharma, and C. Thakur, "Colon cancer classification of histopathological images using data augmentation," in *Proc. Int. Conf. Control, Autom., Power Signal Process. (CAPS)*, Dec. 2021, pp. 1–5.
- [14] S. Garg and S. Garg, "Prediction of lung and colon cancer through analysis of histopathological images by utilizing pre-trained CNN models with visualization of class activation and saliency maps," in *Proc. 3rd Artif. Intell. Cloud Comput. Conf.*, Dec. 2020, pp. 38–45.
- [15] K. Adu, Y. Yu, J. Cai, K. Owusu-Agyemang, B. A. Twumasi, and X. Wang, "DHS-CapsNet: Dual horizontal squash capsule networks for lung and colon cancer classification from whole slide histopathological images," *Int. J. Imag. Syst. Technol.*, vol. 31, no. 4, pp. 2075–2092, Dec. 2021.
- [16] Y. Su, X. Tian, R. Gao, W. Guo, C. Chen, C. Chen, D. Jia, H. Li, and X. Lv, "Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105409.
- [17] A. S. Sakr, N. F. Soliman, M. S. Al-Gaashani, P. Pławiak, A. A. Ateya, and M. Hammad, "An efficient deep learning approach for colon cancer detection," *Appl. Sci.*, vol. 12, no. 17, p. 8450, Aug. 2022.
- [18] Y. G. C. Kelana, S. Rizal, and S. Saidah, "Classification of histopathological images of colon cancer using convolutional neural network method," in *Proc. Int. Conf. Comput. Sci., Inf. Technol. Eng. (ICCoSITE)*, Feb. 2023, pp. 821–826.
- [19] R. D. Mohalder, F. Bin Ali, L. Paul, and K. Hasan Talukder, "Deep learning-based colon cancer tumor prediction using histopathological images," in *Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2022, pp. 629–634.
- [20] A. Bhattacharya, B. Saha, S. Chattopadhyay, and R. Sarkar, "Deep feature selection using adaptive β -Hill climbing aided whale optimization algorithm for lung and colon cancer detection," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104692.
- [21] I. D. Irawati, I. Andrea Larasaty, and S. Hadiyoso, "Comparison of convolution neural network architecture for colon cancer classification," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 18, no. 3, pp. 164–172, Mar. 2022.
- [22] S. S. R. Bairaboina and S. R. Battula, "Ghost-ResNeXt: An effective deep learning based on mature and immature WBC classification," *Appl. Sci.*, vol. 13, no. 6, p. 4054, Mar. 2023.
- [23] S. Sorguli and H. Rjoub, "A novel energy accounting model using fuzzy restricted Boltzmann machine—Recurrent neural network," *Energies*, vol. 16, no. 6, p. 2844, Mar. 2023.
- [24] Y. Li and Y. Li, "Predicting chaotic time series and replicating chaotic attractors based on two novel echo state network models," *Neurocomputing*, vol. 491, pp. 321–332, Jun. 2022.
- [25] S.-M. Guo, J.-K. Guo, Y.-G. Gao, P.-Y. Guo, F.-J.-A. Huang, S.-C. Wang, Z.-C. Lou, and X. Zhang, "Research on engine speed control based on tuna swarm optimization," *J. Eng. Res. Rep.*, pp. 272–280, Dec. 2022.
- [26] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and Colon Cancer Histopathological Image Dataset (LC25000). Accessed: Mar. 14, 2023. [Online]. Available: <https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancerhistopathological-images?resource=download>
- [27] H. A. Mengash, M. Alameer, M. Maashi, M. Othman, M. A. Hamza, S. S. Ibrahim, A. S. Zamani, and I. Yaseen, "Leveraging marine predators algorithm with deep learning for lung and colon cancer diagnosis," *Cancers*, vol.