

# Advances in Artificial Intelligence for Lung Cancer Detection and Diagnostic Accuracy: A Comprehensive Review

Rupa Debnath<sup>1</sup>; Rituparna Mondal<sup>2</sup>; Arpita Chakraborty<sup>3</sup>; Siddhartha Chatterjee<sup>4</sup>

<sup>1</sup>Department of Computational Sciences, Brainware University, West Bengal-700125, India

<sup>2</sup>Department of Computer Applications, Techno India University, Kolkata-700091, West Bengal, India

<sup>3</sup>Department of Computer Science and Engineering, Techno Bengal Institute of Technology, Kolkata-700150, West Bengal, India <sup>4</sup>Department of Computer Science and Engineering, College of Engineering and Management Kolaghat, Purba Medinipur – 721171, West Bengal, India

Publication Date: 2025/05/26

**Abstract:** Cancer is a deadly disease with a minimal probability of curing when detected in the later stages. Out of many different varieties of cancer diseases, lung cancer falls under the critical category as it is internal, invisible, and connected to the breathing control mechanism of human beings. Lung cancer has become the most frequent type in this generation and has intensified in the post-COVID-19 pandemic era, with a high degree of fatality. The histopathological images of lung cancer are very large in volume, so that analyzing such kind of data is monotonous and error-prone in a human-controlled method. The recent progress in the field of computer vision, along with machine learning techniques, has made the path of research smooth in the medical and healthcare domain, also achieved the feasibility of data analysis with easy detection of cancer cells. The advent of the deep learning concept made the automatic and accurate detection of cancer cells from the histopathological image data analysis possible. In this paper, an evaluation and a methodological survey on Cancer cell detection and the accuracy benchmark of Cancer tissue Segmentation of whole slide images (WSI) using Machine Learning (ML) and Deep Learning (DL) approaches have been carried out. The critical analysis, along with the exploration of more probable research trends in the accurate interpretation of the cancer cell images, has also been addressed towards achieving the greater potential.

**Keywords:** Lung Cancer, Adenocarcinoma, Whole Slide Images, Convolution Neural Network, Generative Adversarial Network, Class Activation Mapping, Artificial Intelligence, Deep Learning, Image Processing.

**How to Cite:** Rupa Debnath; Rituparna Mondal; Arpita Chakraborty; Siddhartha Chatterjee (2025) Advances in Artificial Intelligence for Lung Cancer Detection and Diagnostic Accuracy: A Comprehensive Review.

*International Journal of Innovative Science and Research Technology*, 10(5), 1579-1586.

<https://doi.org/10.38124/IJISRT/25may1339>

## I. INTRODUCTION

In the leading causes of death in this era, cancer is one of the most frightening and threatening names of diseases that leads people to death or a costly medication-based survival with no freedom of livelihood. Compared to all the other cancer types, lung cancer is one of the leading cancers across the world. According to WCRF, lung cancer is the second most common cause of cancer in recent times across the globe. This cancer is basically of two different cell types; it may be small cell lung cancer (SCLC) or it can be non-small cell lung cancer (NSCLC) conventionally. So, as per the symptoms, the cancer cells need to be detected in the early stages with immunotherapy and infection-detection targeted therapy[1].

The emerging clinical data analysis of digital pathology and radiology probabilities of developing, as well as application-based experiments of the histopathological dataset, has made the clinical data analysis more accurate. During the 2000 to 2016 study of Germany, 85,685, that is 46.1%, of females and 100,282, which is 53.9%, were males out of 185,967 patients diagnosed with colon cancer cells, which is not negligible[2]. However, the accuracy benchmark needs to be re-analysed as there are risk factors of inaccurate diagnosis and decision support systems. The detection of cancer cells is performed through the histopathological image analysis. The image processing algorithms, along with the machine learning models, have produced very accurate and good responses. In this paper, an in-depth review has been carried out to assess the accuracy

obtained in different research works comparatively, based on the fundamental principles of machine learning techniques and their extension in the deep learning paradigms.

The paper is organized as follows: Section II provides the literature survey, where Section II.A shows an overview of the most frequently and relatively used datasets and the anatomy of the employed histopathological image and WSI slides, where some are collected sets of data and some are public datasets. The section II.B elucidates the related work part that is some of the research on cancer cell classification, the accuracy analysis and Area Under the Curve (AUC) using Machine Learning followed by Deep Learning Methodologies, Convolution Neural Network (CNN) and finally section III briefs about the future scope of work on the same and the entire study conclusion.

## II. LITERATURE STUDY

### A. Anatomy of the Applied Datasets

To analyse-adenocarcinomas in detail, it can be said that it is a kind of malignant tumour having differentiation concerning epithelial cells, which also contains clear patterns of morphology like papillary, acinar, micro-papillary, and lepidic, where the mixed pattern of

adenocarcinomas is one of the most common patterns found[3][4]. Research activity in this area of cancer cell identification has been traced back almost to the middle of the last century, along with the application of image data analytics to a particular set of medical images of human cells. The algorithms can classify the images of animal cells or tissue based on the characteristics of the quantity, that is, the chromatin distribution, shape, size, and the support diagnosis of diseases. The algorithms used in the implementation of health care or medical image analysis were later replaced by Machine learning based radiographic images using texture-describing features[5].

#### ➤ LC25000 Dataset:

The LC25000 lung and colon histopathological image dataset contains 25000 colour images of 5 different classes, with each of the classes having 5000 ready images of 768 x 768 pixels in size and a JPEG file extension. The Lung image sets subfolder contains two more secondary subfolders where one is named Lung aca subfolder having 5000 images of lung adenocarcinomas and the other subfolder named lung n contains another 5000 images of benign lung tissues and lung squamous cell carcinomas tissues, one more subfolder colon image sets also having two subfolders called colon aca and colon n with 5000 images of colon adenocarcinomas and another 5000 images of benign colonic tissues respectively[6].

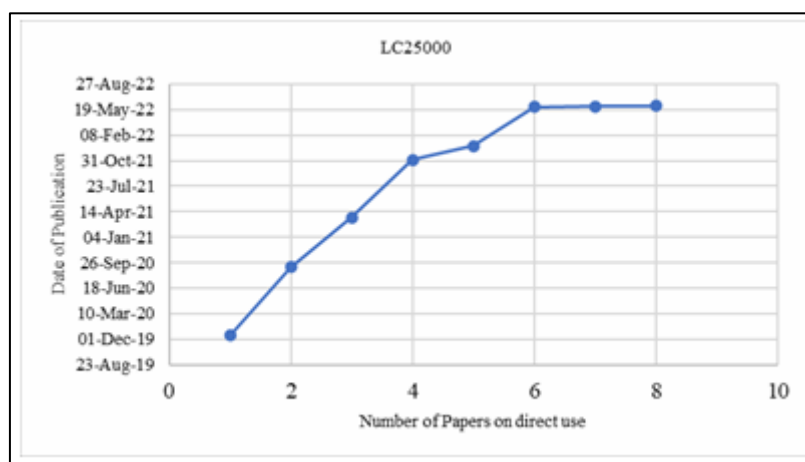


Fig. 1. Data Visualization of the Papers Published from 2019 to 2022 Using the LC25000 Dataset.

Fig. 1 shows the statistical plotting of the number of papers directly used over the years, and it is evident from the graph. In the three big challenges, like detecting the cancer cell, segmentation of related images, and classifying the nuclei [7] [8], also for the segmentation of large organ images[9][10] and to classify the affected cell and analyze the lesion with the enhanced large-scale giga pixel slide images availability of the specimen of issue, the application of CNN is subjected to the domain, and also CNN applications are process-specific in normalizing the colour of histopathological images[11][12].

#### ➤ WSI Dataset (TCGA):

The dataset initially consisting of 741 haematoxylin and LUAD, LUSC, and SCLC Whole Slide Images Slides (WSI) dataset contains 421 WSI of colon tumors associated

with the generic mutation of colon cancer profiles from Cbio-portal, which is collected from Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD)[13]. In the anatomy of the slides, the features of trained models and the learned representation get affected by normalization and augmentation strategies; as a result, the training data gets influenced heavily during the preparation of the training data[14].

#### ➤ NCT Dataset:

The “NCT-CRC-HE-100K” dataset contains 100,000 non- overlapping H&E stained histological image sets of normal human tissue also human colorectal cancer (CRC) images where 5000 of the images are of colorectal cancer and the rest of the dataset is comprised with nine organizational categories where with external validation of

7180 images; to be mentioned all the images are of 224x224 pixels at 0.5 MPP also The large-scale experimental validation is done based on large cohorts of the 100,000 histological image dataset by TCGA[15][16]. For accuracy analysis using CNN, the NCT-CRC-HE-100K dataset is widely used in the expert evolution of healthcare, or to be more specific, in cancer cell detection algorithms, where [17] proposed a unique ensemble network for tracing tumours in [18][19][20] the colorectal histology images, with an achieved accuracy value is approximately 96%.

#### ➤ Lung\_Cancer Dataset:

Thanks to the effectiveness of cancer prediction systems, People can learn their cancer risk at a reasonable cost and make the right decision based on their cancer risk status [21]. The online lung cancer prediction system is the source of the data. There are sixteen qualities in total. There are 284 examples. Information about attributes:

#### B. Related Work

In certain groups and researchers, deep learning-based frameworks and computer vision have become a choice for medical data analysis. In the evaluations of tumour histopathology, the intensive tasks such as segmentation, detection of the region of interest for the

element quantification, and visualization of the histopathology have led to a good execution of deep learning algorithms [18][22] Besides, a lot of deep learning has led to the successful classification of lung cancer histopathological slides under the supervision of the models of CNN [22][23].

The following are some of the healthcare domains where the use of elite Artificial Intelligence algorithms can be seen in various studies:

- The algorithm that finds differences in a tumor.
- Heart image classification.
- AI algorithms to predict cardiovascular diseases.
- More accurate diagnosis of skin cancer using AI
- AI that predicts the risk of breast cancer.
- AI that predicts the risk of colon cancer.
- Histopathological image analysis to identify cancer cells by using image processing.

There are some research results analysis along with accuracy, the top AUC, and the tools for implementation used have been aligned in the Table. 1 and Table. 2, respectively.

TABLE I. RELATED WORK WITH RESULT ANALYSIS

<i>References</i>	<i>Results</i>
[4]	AUC were 0.733 to 0.856
[20]	Cancerous tissue with 96.38% of F-measure score
[23]	Cancerous cell identification accuracy rate 99.7%
[22]	Competitive accuracy 99.87%
[24]	MA ColonNET accuracy 99.75%
[25]	Classification accuracy was 83.3%
[12]	The mean PTS Score was 0.380
[13]	Mean DICE coefficient 0.7966 and 0.7544

Table. 1 discloses that Deep learning based models are used for the segmentation of the histopathological images in the sparsely annotated dataset, and the Performance values.

TABLE II. RELATED WORK WITH RESULT ANALYSIS

<i>References</i>	<i>Top AUC</i>	<i>Accuracy</i>	<i>Tools and Algorithms Used</i>
[4]	0.856		DCNN (inception v3)
[20]		96.38%	CNN (Model for classification)
[23]		99.7%	ResNet 50, VGG-19, DenseNet
[22]		99.87%	Vision Trasformer Model
[24]		99.75%	MATLAB 2019b
[25]		83.3%	MATLAB

Table. 2 shows the variation of accuracy values, with AUC plotting being evident. The highest accuracy achieved has been found as 99.87%. The accuracy value is closest to 100%, which confirms the potential of a deep learning-based model in the detection of cancer cells. Some of the notable computational tools like DeepFocus, ACD model, QuPath, ConvPath and HistQC[24][25][26] have been developed and implemented to perform

annotating, data mining, viewing, and whole slide imaging (WSI) for analysing WSI analysing tools[27]. The workflow of the study for deep learning principles for six types of classifiers based on the efficient threshold methods for slide label inference [28][29].

### III. METHODOLOGIES AND ALGORITHMS USED

Machine learning and deep learning algorithms have generated innovations for years in the healthcare sector. Decision-making support, enhanced image scanning and segmentation, case triage and diagnosis, disease risk

prediction, and neuroimaging are all possible with AI [30][31][32]. Machine learning aims to develop algorithms that can evaluate and interpret data, identify trends, and make well-informed decisions or predictions. It includes many different methods and strategies, such as reinforcement learning, supervised learning, and unsupervised learning[33][34].

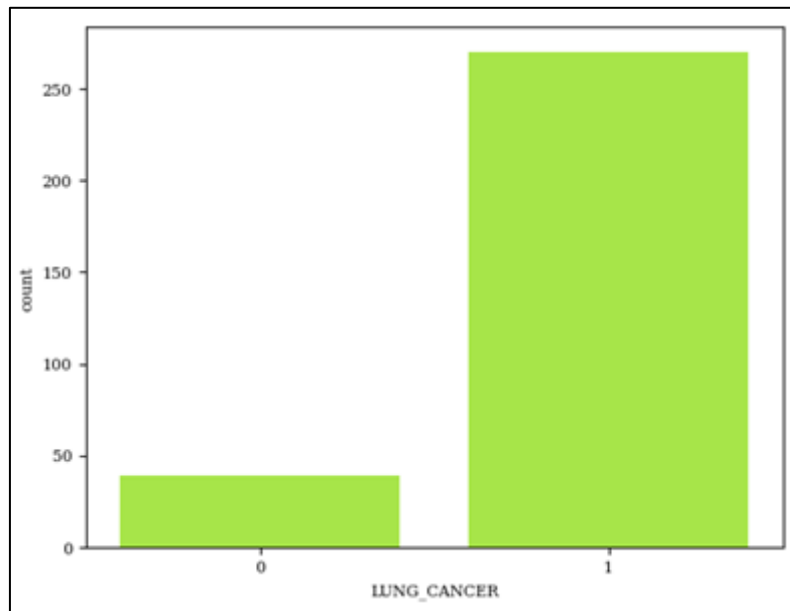


Fig. 2. Data Visualization of the Count of Patients with Lung Cancer out of 309 Patients of the Lung-Cancer Public Dataset.

Here, Fig.2 shows the number of cases detected in the public dataset called “Lung Cancer”. The early detection of lung cancer is the main emphasis of this investigation. The non-parametric Genetic K-Nearest Neighbour (GKNN) algorithm is suggested for the detection in some studies, also to overcome the time-consuming and crucial nature of manually interpreting lung cancer CT pictures. The Genetic Algorithm approach is integrated with the K-Nearest Neighbour (K-NN) algorithm, which will classify the cancer images efficiently[35].

Analysis is done on performance metrics such as false positive rates and classification rates. The distance between each test and training sample is first determined in the classic Naive Bayes, SVM, Decision Tree, and Logistic Regression algorithms. Each iteration of the employed techniques selects a certain number of samples, and fitness is defined as a 90% classification accuracy. Every time, the highest accuracy is noted[36]. In the medical field, artificial neural networks (ANN) have been used for lung cancer analysis and prognosis. This paper discusses current lung cancer risk factors and the use of machine learning algorithms, paying particular emphasis to their relative advantages and disadvantages as well[37].

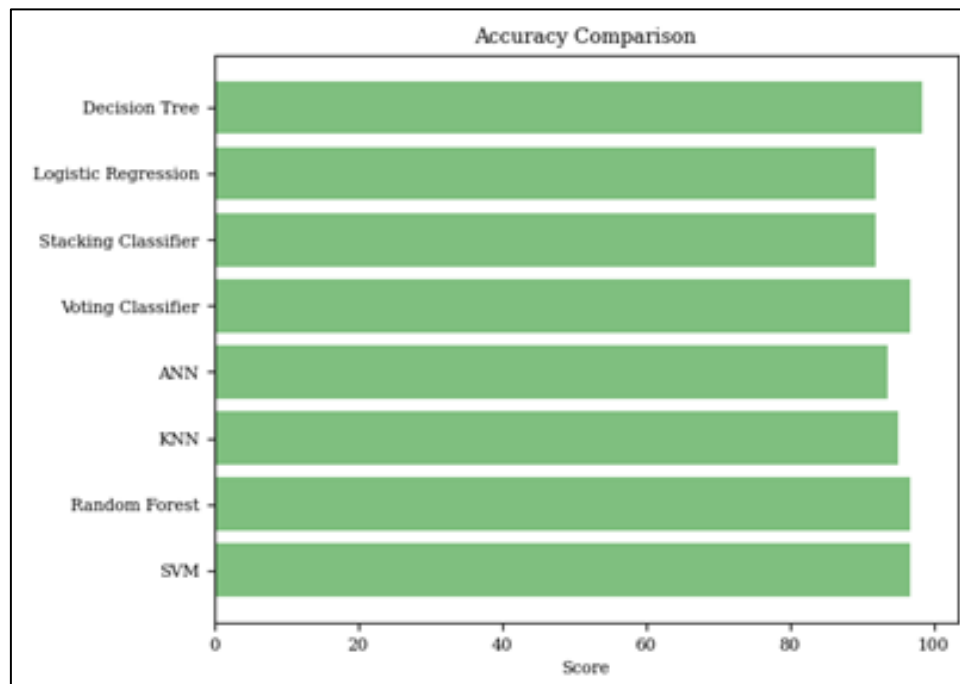


Fig. 3. Data Visualization of the Machine Learning Algorithms Used with the Lung-Cancer Public Dataset and its Accuracy Benchmark.

Fig. 3. The above plot shows the machine learning algorithms used to check the accuracy benchmarking of the different algorithms, and whichever suits the purpose and provides relevant accuracy in different studies. As per the study, we can see that Support Vector Machine (SVM), Random Forest, Voting Classifier, and Decision Tree are some of the competent machine learning algorithms.

A glance at comparisons between the most frequently used ML algorithms can be discussed, such as SVM is the most widely used technique for classification, regression, and prediction. By creating a boundary known as a hyperplane that divides the incoming dataset into two parts, it classifies the dataset, whereas an ANN is a computing technique that is severely hampered by interconnected processing units called neurons that categorize data as feedback to external stimuli. There are two ways to do it, like learning and testing.

New input is categorized by learning. In the attesting phase, it computes an input signal from the network and produces an output.

#### IV. CONCLUSION AND FUTURE WORK DIRECTION

Lung cancer detection using computer vision and machine learning concepts has revolutionised the traditional treatment procedure to a high degree, as greater accuracy has been achieved with higher precision. The imperfections in the image make the process unfit for accurate detection of the infected cell or tumour. The advent of machine Learning and deep learning thoughts and their application capability made the roadways to the precision learning and accurate interpretation of the image much smoother, and hence achieved a greater success rate. In this review paper, many deepening-based algorithms have been discussed. Most of such algorithms work in a non-invasive manner. Machine learning procedure does not demand any invasive way of detection of the cancerous cells or tissues. In this review paper, many deepening-based algorithms have been discussed. Most of such algorithms work in a non-invasive manner.



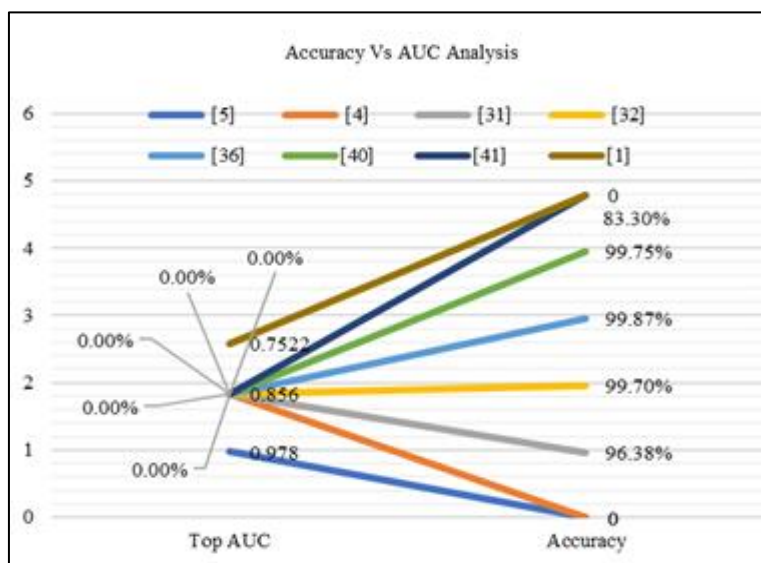


Fig. 4. Data visualization of the Accuracy Vs Area Under the Curve analysis.

Fig. 4. shows the comparison between the top AUC parameter and the result accuracies of different studies using ML and DL algorithms.

This paper dealt with several machine learning algorithms, such as SVM, Random Forest, Voting Classifier, Decision Tree, Logistic Regression, Stacking Classifier, as well as deep learning methodologies, like ANN, CNN, KNN, RNN, RBFN. Despite individual shortcomings of each of the algorithms, the CNN in deep learning algorithms and SVM, Random Forest, and Decision Tree in machine learning algorithms provided a more accurate result of classification due to their proximity to the domain of computer vision. It is obvious from the analysis of the review papers that most of the research works are concerned with the detection of lung cancer from the image of the cancerous cells. The degree of infection of the cell cannot be answered well from the research. The future trend of the research should therefore be extended to find the state of the cancer infection and the corresponding stages of infection.

The concept of auto-organisation has been considered as an emerging field. This process refers to unsupervised learning. The main objective is to find the features and relations among different patterns existing inside the features. The author helps to increase the number of levels of feature representation[38]. The trend is still under research, and it also contributes to the improvement of the accuracy of the system.

## REFERENCES

- [1]. Y. Zheng, D. Liu, B. Georgescu, D. Xu, and D. Comaniciu, "Deep learning based automatic segmentation of pathological kidney in CT: local versus global image context," in *Deep learning and convolutional neural networks for medical image computing*, Springer, pp. 241–255, 2017.
- [2]. R. Schmuck et al., "Gender comparison of clinical, histopathological, therapeutic and outcome factors in 185,967 colon cancer patients," *Langenbecks Arch Surg*, vol. 405, no. 1, pp. 71–80, Feb. 2020.
- [3]. W. D. Travis et al., "The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification," *Journal of Thoracic Oncology*, vol. 10, no. 9, pp. 1243–1260, Sep. 2015.
- [4]. J. Kim and J. D. Minna, "AP-1 leads the way in lung cancer transformation," *Dev Cell*, vol. 57, no. 3, pp. 292–294, 2022.
- [5]. [5] A. Zwanenburg et al., "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.
- [6]. 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)," Dec. 2019, Accessed: May 15, 2025.
- [7]. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med Image Anal*, vol. 42, pp. 60–88, Dec. 2017.
- [8]. X. Pan et al., "Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks," *Neurocomputing*, vol. 229, pp. 88–99, 2017.
- [9]. F. Mahmood et al., "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans Med Imaging*, vol. 39, no. 11, pp. 3257–3267, 2019.
- [10]. S. Kannan et al., "Segmentation of Glomeruli Within Trichrome Images Using Deep Learning," *Kidney Int Rep*, vol. 4, no. 7, pp. 955–962, Jul. 2019.
- [11]. A. Patil et al., "Fast, Self Supervised, Fully Convolutional Color Normalization Of H&E Stained Images," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, pp. 1563–1567, 2021.

- [12]. D. Ziaei et al., "Characterization of color normalization methods in digital pathology whole slide images," *SPIE*, vol. 11320, p. 1132017, Mar. 2020.
- [13]. K. Ding, Q. Liu, E. Lee, M. Zhou, A. Lu, and S. Zhang, "Feature-Enhanced Graph Networks for Genetic Mutational Prediction Using Histopathological Images in Colon Cancer," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12262 LNCS, pp. 294–304, 2020.
- [14]. K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "A closer look at domain shift for deep learning in histopathology," *arXiv preprint arXiv:1909.11575*, 2019.
- [15]. M.-J. Tsai, Y.-H. Tao Citation, Y. Deep, and A. Wysocki, "Deep Learning Techniques for the Classification of Colorectal Cancer Tissue," *Electronics* 2021, Vol. 10, Page 1662, vol. 10, no. 14, p. 1662, Jul. 2021.
- [16]. Y. Shen and J. Ke, "Su-Sampling Based Active Learning For Large-Scale Histopathology Image," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 116–120, 2021.
- [17]. S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. C. Santosh, "Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network," *Eng Appl Artif Intell*, vol. 100, Apr. 2021.
- [18]. H. H. N. Pham, M. Futakuchi, A. Bychkov, T. Furukawa, K. Kuroda, and J. Fukuoka, "Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach," *Am J Pathol*, vol. 189, no. 12, pp. 2428–2439, 2019.
- [19]. K. H. Yu et al., "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 757–769, May 2020.
- [20]. M. Yildirim and A. Cinar, "Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN model: MA\_ColonNET," *Int J Imaging Syst Technol*, vol. 32, no. 1, pp. 155–162, Jan. 2022.
- [21]. S. Hazra et al., "Pervasive Nature of AI in the Health Care Industry: High-Performance Medicine," *International Journal of Research and Analysis in Science and Engineering*, vol. 4, no. 1, pp. 16–16, Jan. 2024, Accessed: May 18, 2025.
- [22]. K. H. Yu et al., "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 757–769, May 2020.
- [23]. M. Yildirim and A. Cinar, "Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN model: MA\_ColonNET," *Int J Imaging Syst Technol*, vol. 32, no. 1, pp. 155–162, Jan. 2022.
- [24]. A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, "HistoQC: an open-source quality control tool for digital pathology slides," *JCO Clin Cancer Inform*, vol. 3, pp. 1–7, 2019.
- [25]. M. S. Kwak et al., "Deep Convolutional Neural Network-Based Lymph Node Metastasis Prediction for Colon Cancer Using Histopathological Images," *Front Oncol*, vol. 10, Jan. 2021.
- [26]. A. Das, M. S. Nair, and S. D. Peter, "Computer-Aided Histopathological Image Analysis Techniques for Automated Nuclear Atypia Scoring of Breast Cancer: a Review," vol. 33, no. 5. *Journal of Digital Imaging*, 2020.
- [27]. A. Ben Hamida et al., "Deep learning for colon cancer histopathological images analysis," *Comput Biol Med*, vol. 136, no. July, 2021.
- [28]. Z. Li et al., "Deep Learning Methods for Lung Cancer Segmentation in Whole-Slide Histopathology Images - The ACDC@LungHP Challenge 2019," *IEEE J Biomed Health Inform*, vol. 25, no. 2, pp. 429–440, Feb. 2021.
- [29]. P. Ghosh and S. Chatterjee, "Future Prospects Analysis in Healthcare Management Using Machine Learning Algorithms," *the International Journal of Engineering and Science Invention (IJESI)*, ISSN 2319-6734, 2023.
- [30]. A. Gon, S. Hazra, S. Chatterjee, and A. K. Ghosh, "Application of Machine Learning Algorithms for Automatic Detection of Risk in Heart Disease," *Cognitive Cardiac Rehabilitation Using IoT and AI Tools*, pp. 166–188, Jan. 2023.
- [31]. S. Das, S. Chatterjee, D. Sarkar, and S. Dutta, "Comparison Based Analysis and Prediction for Earlier Detection of Breast Cancer Using Different Supervised ML Approach," pp. 255–267, 2023.
- [32]. S. Das, S. Chatterjee, A. I. Karani, and A. K. Ghosh, "Stress Detection While Doing Exam Using EEG with Machine Learning Techniques," pp. 177–187, 2024.
- [33]. A. Adhikary, S. Das, R. Mondal, and S. Chatterjee, "Identification of Parkinson's Disease Based on Machine Learning Classifiers," pp. 490–503, 2024.
- [34]. P. Ghosh, R. Dutta, N. Agarwal, S. Chatterjee, and S. Mitra, "Social Media Sentiment Analysis on Third Booster Dosage for COVID-19 Vaccination: A Holistic Machine Learning Approach," *Lecture Notes in Electrical Engineering*, vol. 985, pp. 179–190, 2023.
- [35]. E. Ajitha, B. Diwan, and M. Roshini, "Lung Cancer Prediction using Extended KNN Algorithm," *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, pp. 1665–1670, 2022.
- [36]. P. R. Radhika, R. A. S. Nair, and G. Veena, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*, Feb. 2019.

- [37]. S. S. Raoof, M. A. Jabbar, and S. A. Fathima, “Lung Cancer Prediction using Machine Learning: A Comprehensive Approach,” 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 Conference Proceedings, pp. 108–115, Mar. 2020.
- [38]. R. Chatterjee, S. Chatterjee, S. Samanta, and S. Biswas, “AI Approaches to Investigate EEG Signal Classification for Cognitive Performance Assessment,” Proceedings - International Conference on Computational Intelligence and Networks, 2024.