Vision Transformer based Automated Model for Enhancing Lung Cancer Classification

Akbar Sheikh Akbari
Leeds Beckett University, UK
A.Sheikh-Akbari@leedsbeckett.ac.uk

Arvind Kumar
Machine Vision and Intelligence Lab,
National Institute of Technology
Jamshedpur, Jharkhand,India
arvindkumarmm02@gmail.com

B Ramachandra Reddy
Computer Science and Engineering
National Institute of Technology
Jamshedpur, Jharkhand, India
brcreddy.cse@nitjsr.ac.in

Koushlendra Kumar Singh
Machine Vision and Intelligence Lab,
National Institute of Technology
Jamshedpur, Jharkhand, India

Masahiro TAKEI
Department of Mechanical Engineering,
Chiba University, Chiba, Japan,
masa2@chiba-u.jp

***Abstract* –Lung cancer is one of the leading causes of cancer related mortality. The early detection and classification of the cancers tissues will reduce the mortalities rate. The present research focus on the development of automated classification model for lung and colon cancers tissues based on the histopathology images. The present work encompasses a vision transformer (ViT) based model to enhance diagnostic accuracy of lung cancers tissues. The proposed model utilizes the self-attention mechanism of ViT to focus on essential features present in histopathologicals images. The proposed model has been validated using two different dataset namely LC25000 & IQ-OTH/NCCD with 25000 & 1096 images respectively. The performance of proposed model is compared with traditional convolutional neural network (CNN) model and it has been observed the based model outforms better in terms of accuracy which – 98.80% & 99.09% respectively for datasets.**

Keywords - ViT model, ROC-AUC curve, CMC curve.

I. INTRODUCTION

The unchecked growth of aberrant cells in lungs is known as lung cancer. Cigarettes are the leading cause of the majority of the lung cancer and colon cancer cases. Lung cancer is the worst malignancy, taking the lives of 1.8 million people per year with a fatality rate ranging from 80 to 90 %, it is the primary cause of cancer-related deaths in men and the second most common in women [2]. The great majority of people who have lung cancer have smoked for a long time. On the other hand, some forms of lung cancer can strike in the healthy people who have never smoked. Small Cell and Non Small-Cell Lung Cancer are the two primary forms of lung cancer. While Non Small Cell Lung Cancer is typically treated with surgery to remove the lung tumour, radiotherapy is also an option. Roughly 85-90% of instances of lung cancer are Non Small Cell Lung Cancer. An increased risk of lung cancer exists in those with a markedly decreased pigment count. Lung cancer, which is not tiny cells, exists in three varieties. Squamous cell, Adenocarcinoma and Large cell carcinoma

The paper's overall structure is set out as follows: The suggested work is introduced in Section I. Some more recent work that is relevant to the current work is reported in Section II. The dataset and has been methodology used in the proposed work are described in Section III.

II. LITERATURE REVIEW

The application of AI in diagnosis of lung and colon cancer is still interesting areas for researcher of artificial intelligence (AI) and medical imaging Traditional AI based algorithms as well as ensemble models have been extensively used to detect the lung and colon cancer. It has been observed that ensemble classifier composed of logistic regression, support vector machines and a random forest has been regressively used to detect the cancers tissues. Deep learning based model becomes quite popular in automatic diagnosis of cancer disease. The findings demonstrate that ensemble of CNN model with traditional machine learning (ML) model algorithms technique outperforms in terms of the accuracy of cancer detection prediction [12]. Currently vision transformers have been became increasingly helpful for medical imaging application. Mehta et al, Explored vision transformer in the field of biometric for ear recognition. Vision Transformers (ViTs) have become increasingly helpful for machine translation and image recognition tasks. Its enhanced efficiency over current models has been proven on the IITD-II [13]. Ruina, The Swin Transformer has been used for segmenting and classification of lung cancer to assist radiologists in making decision. Pre-trained Swin-B and swing-S model outperformed ViT by 2.529% with a top-1 classification accuracy of 82.26% and in terms of Mean Intersection over Union (MIoU). These findings demonstrate that pre-training increases the accuracy of the Swin Transformer in these tasks [14]. The classification of lung cancer subtypes from histopathological imaged in the LC25000 dataset using pre-trained Vision Transformer (ViT) model has been used. ViT performs well even with a small amount of training data; after one round in the Few-Shot mode, it achieves a comparative accuracy of 99.87% and after five rounds, it reaches its ideal accuracy of 100%. A comparison between Zero-Shot and Few-Shot ViT demonstrates how well it can classify lung cancer [15]. Pathologists often use a laborious process of histopathologicals image analysis to diagnose hepatocellular carcinoma (HCC), a deadly disease with a high death rate [6]. To help pathologists automatically assess HCC pictures, this research offers a dual-channel attention-sharing hybrid network (DCAH-Net) that uses multi-feature fusion. This study validated on the LC25000 and Break His datasets, DCAH-Net had significant potential for various medical classification tasks and attained the highest performance metrics on the HCC dataset, including

accuracy (88.36%), macro recall (82.14%), macro precision (83.00%), and macro F1-score (82.30%) [16]. Han et al. Proposed a model to diagnose cancer from histopathology images, introduces a feature learning enhanced convolutional neural network (FLE-CNN). The FLE-CNN has a residual fusion unit and an information refinement unit with dual-domain attention to extract incredibly detailed features. The model outperforms previous sophisticated models with Average Sensitivity, Specificity, Accuracy, Precision, and F1-Scores of 0.9992, 0.9998, 0.9992, 0.9997, and 0.9992, respectively [17]. Chillars et al. has been proposed histopathology images to offer a simple machine-learning strategy for classifying canceres tumors. A Light Gradient Boosting Machine (LightGBM) classifies the features after merging texture characteristics extracted with Haralick and colour features extracted with a colour histogram. LightGBM obtained 97.72% accuracy with texture features, 99.92 % with colour characteristics, and 100% accuracy with combined features on the LC25000 dataset [18]. Trishna explored the lung cancer an automatic hybrid and transfer learning based classification of cancers using a VGG and random forest ,support vector machine used to achieved highest accuracy 98.70% [20].The present work contributes in literature in terms of developing a vision transformer based model to classify a lung and colon cancer using image patches instead of different features. The proposed model has been validated to LC25000 and IQ-OTH/NCCD it has been observed that proposed that out performs with existing method in terms of accuracy.

Table 1 Literature review of the proposed work.

Authers	***Dataset***	***Classification***	***Accuracy***
Onkar [12]	LC25000	Ensembled	99.00%
Mehta et al [13]	IIT D	ViT	99.36%
Ruina [14]	Histopathological	Swin transformer	82..26%
Fu-Ming [15]	LC25000	ViT	99.87%
Zhang [16]	LC25000	DCAH-Net	88.36%
Han et al [17]	Histopathological	FLE-CNN	99.96%
Chhillar [18]	Histopathology	Light -GBM	99.92%
Trishna [20]	CT image	VGG,SVM,RF	98.70%
Proposed*	LC25000, IQ-OTH/NCCD	ViT	99.06%

III. Materials and Methodology

A. Model Architecture

The Vision Transformer (ViT) model is a pioneering architecture in computer vision that revolutionises traditional approaches by leveraging transformer networks originally designed for natural language processing. ViT operates by first breaking down input images into fixed-size patches, which are then linearly embedded along with learnable positional encodings. These embeddings are processed through multiple transformer encoder layers, where self-attention mechanisms capture global dependencies among image patches, enabling long-range interactions. Following the transformer layers, feedforward networks independently process each token, culminating in a classification head that outputs class probabilities. ViT's parameters are optimised during training using backpropagation and gradient-based optimisation algorithms. Through fine-tuning and transfer learning, ViT demonstrates robust performance across various computer vision tasks, offering scalability to larger image resolutions and interpretability. Vision transformer is a very lightweight and effective model. We have used it in the large datasets and a long range of dependencies of image input like LC25000 and IQ-OTH/NCCD dataset. The vision transformer is formally developed for natural language processing tasks.

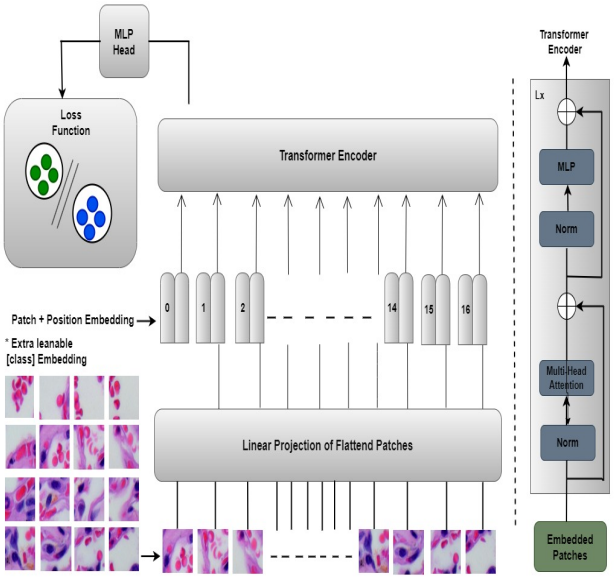

Figure 1 Proposed Vision Tranformer based model architecture with each step in the proposed model

Our lung cancer detection framework is built upon the Vision Transformer (ViT) architecture, which has shown remarkable success in various computer vision tasks. The ViT model comprises a sequence to sequence transformer encoder consisting of Multi Layer Perceptron (MLP) blocks and multi-head self-attention mechanisms. Each transformer block is augmented with layer normalisation to stabilise training and improve convergence. As given below, eq. 1 shows a multilayer perceptron equation.

$$MLP(X) = FC\,(\sigma\,(FC(X))),\ and\ FC(X) = XW + b \quad (1)$$

Multi-head self-attention is a mechanism that can store information on different parts of images of each patch. Each patch has fixed square size images of 16×16 patch size.

B. Dataset Description

The LC25000 and the IQ-OTH/NCCD datasets are two large sets of medical images that make data available for developing and validating lung cancer. At the LC25000 has five classes of 25000 images of high resolution and their IDs are colon adenocarcinoma (ID,0), Colon benign tissue(ID,1), Lung adenocarcinoma(ID,2), Lung benign tissue(ID,3) and lung squamous cell carcinoma(ID,4). The IQ-OTH/NCCD dataset has three classes of 1024 images of high resolution and their IDs, Benign cases (ID, 0), Malignant cases (ID, 1) and normal cases (ID,2) that focus on various cancer-related diseases. This dataset is valuable for developing machine-learning models to classify various

types of colon cancer. The dataset is splitting the ratio of 8:2, which means 80 % of the training data and 20% of the validation and testing of both the LC25000 and IQ-OTH/NCCD datasets. The LC25000 and IQ-OTH/NCCD datasets have given detailed descriptions of images and IDs in Table 2 and Table 3. Figure 1 shows the sample images of Caner related to both the LC25000 and IQ-OTH/NCCD datasets. Dataset description is very accurate and very effective to test and validate the vision transformer model for lung cancer.

Table 2 Class ID and cancer types in LC25000 dataset.

Type of Cancer	Class ID	Class Name	No. of Images
colon adenocarcinoma	0	colon_aca	5000
colon begin tissue	1	colon_n	5000
lung adenocarcinoma	2	lung_aca	5000
lung benign tissue	3	lung_n	5000
lung squamous cell carcinoma	4	lung_scc	5000

Table 3 Class name and class ID of IQ-OTH/NCCD dataset.

Class Name	Class ID	No. of Images
bengin cases	0	560
malignant cases	1	416
normal cases	2	120

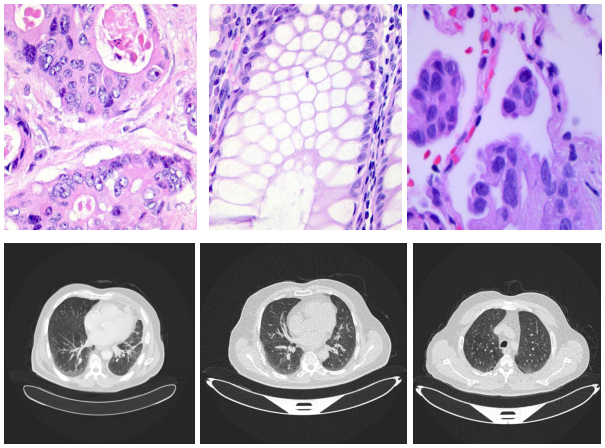

Figure 2 shows the sample images of LC2500 and IQ-OTH/NCCD dataset.

C. CONFIGURING THE HYPERPARAMETER

Configuring Hyperparameters involves selecting the optimal values of parameters that guide the earning process of a machine-learning model but are not updated during training. The common hyperparameter value includes the learning rate, batch size, patch size, number of epochs and regularization terms. Proper tuning is crucial as it directly affects the model's performance, requiring techniques like grid search, random search, or more advanced methods like Bayesian optimization to find the best settings. Firstly, we specify the image size as 224×224 pixels. We have used the Adam optimizer and the categorical cross-entropy in this model. They have utilized the computation of loss function, including model performance and effectiveness. We have used batch size 16 and a learning rate of 0.001 in our model and trained the model with 50 epochs. Table 4 displays the values that we have configured for the hyperparameter.

Table 4 displays the value of the parameter of a dataset.

Hyperparameter	Values
Image Size	224×224
No. of channels	3
Epochs	50
Batch size	32
Activation function	ReLu
Dropout	45%
Dense layer activation	Softmax
Compiler optimizer	Adam
Loss Function	Categorical Cross Entropy
Learning rate	0.001

D. DATA PREPROCESSING

The dataset used for training and evaluation consists of medical imaging data obtained from the LC25000 dataset and the IQ-OTHNCCD datasets of lung cancer. Before training, the input images are resized to a fixed dimension of 224x224 pixels to ensure compatibility with the ViT architecture. Data augmentation methods such as rotation, height and width, shifting, shearing, zooming, and horizontal flipping are applied to augment the training data and improve model generalization.

E. MODEL TRAINING

The ViT model is trained using as an Adam optimiser with the learning rate of 1×10−6. We employ early stopping, model checkpointing, and learning rate reduction on the plateau as callbacks to prevent overfitting and optimise model performance. The model has trained for 50 epochs with the batch size 32, utilisings image data generators for efficient data augmentation and preprocessing.

IV. Results and Discussion

In our experiment, we utilised a Windows 10 computer with an Intel Xeon CPU operating at various speeds and ample RAM. We employed Python and widely-used libraries such as Pandas, NumPy, Matplotlib, TensorFlow, Keras, and Scikit-learn within a Jupyter notebook on Anaconda Navigator. The LC25000 and IQ-OTH/NCCD datasets are split into two groups: 80% for training data and 10% for validation and 10% testing data using the train-test split function. We employed several metrics, such as the ROC curve, accuracy, precision, recall, and F1-score, to assess the effectiveness of our classifier. These metrics show the model's accuracy in differentiating between positive and negative cases. Correctly predicted positive cases are referred to as true positives (TP), correctly predicted negative cases as true negatives (TN), mistakenly forecasted positive cases as false positives (FP), and incorrectly predicted negative cases as false negatives (FN). With an emphasis on colon and lung cancer, we used the Vision Transformer model to classify images into three categories for the IQ-OTH/NCCD dataset and five for the LC25000 dataset. Vision transformers typically need large amounts of training and validation data for optimal performance. We trained the model using our method of training the model using 1,045 photos from the IQ-OTH/NCCD dataset and 25,000 images from the LC25000 dataset. To expand the dataset, we employed techniques for data augmentation,

such as flipping, rotating, and zooming. With this approach, we could effectively train the model and get precise classification results even with a minimal number of source photographs. Table 4 contains a summary of a parameter's tabulation.

Table 5: Training parameter of the ViT model

No. of epochs	*Batch-Size*	*Dataset*	*Image-Size*	*Accuracy (%)*
50	32	LC25000	224	98.80
50	32	IQ-OTH/NCCD	224	99.09

Table 5 shows the adaptability and robustness of our proposed Vision Transformer (ViT) model in different configurations. We optimized the model's performance by modifying the activation function, batch size, patch size, and number of epochs. Following extensive testing, we discovered that using the ReLU activation function, 32 as the patch size, 32 as the batch size, and 50 epochs of the model running yielded terrific results, with an accuracy of 99.08% on the IQ-OTH/NCCD dataset and 98.90% on the LC25000 dataset. Figures 5 and 6 show accuracy and loss graphs for each epoch, demonstrating how well the model works to achieve high accuracy while minimizing loss. This is particularly true if the final layer has a dense layer with softmax activation and patch sizes of 32 and 50 epochs. These results illustrate the flexibility and effectiveness of our ViT model, showing that it can be used for a wide range of vision tasks with the correct parameter selections.

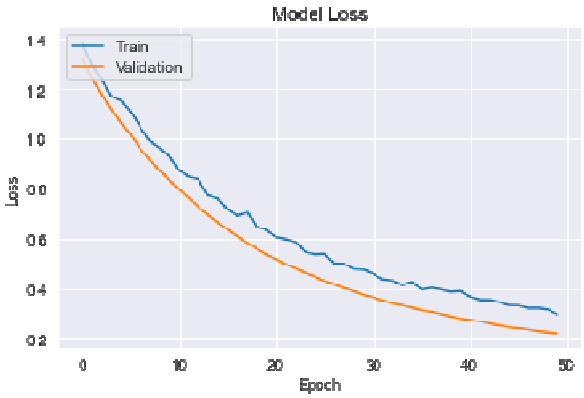

Figure 3 shows the loss curve of the dataset IQ-OTH/NCCD

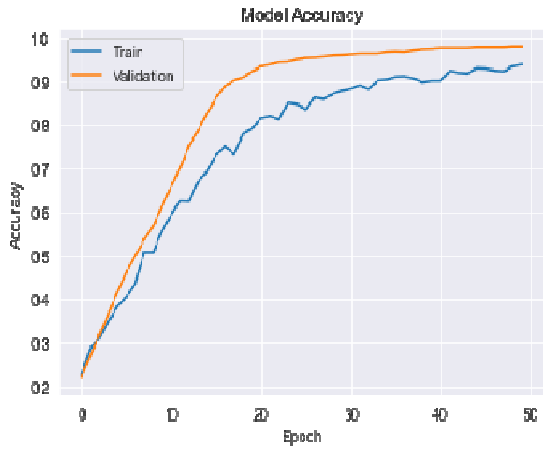

Figure 4 shows the accuracy of the dataset IQ-OTH/NCCD

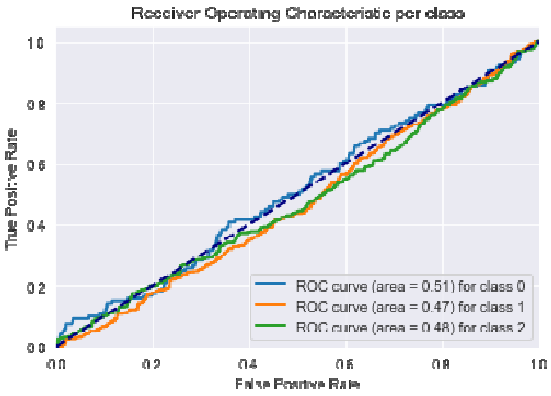

Figure 5 shows the ROC-AUC curve of the dataset IQ-OTH/NCCD

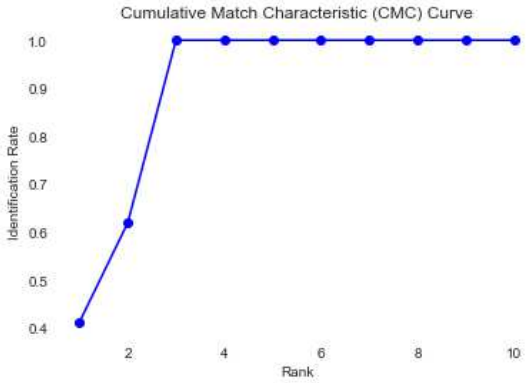

Figure 6 shows the CMC curve of the dataset IQ-OTH/NCCD

The graph of the loss and accuracy curve of the IQ-OTH/NCCD dataset is given above in Figure 3 and Figure 4. After plotting the accuracy and loss curve, calculate the confusion matrices in Figure 7 below, which represent the x-axis indicating the predicted label and the Y-axis representing the true labels within three IQ-OTH/NCCD

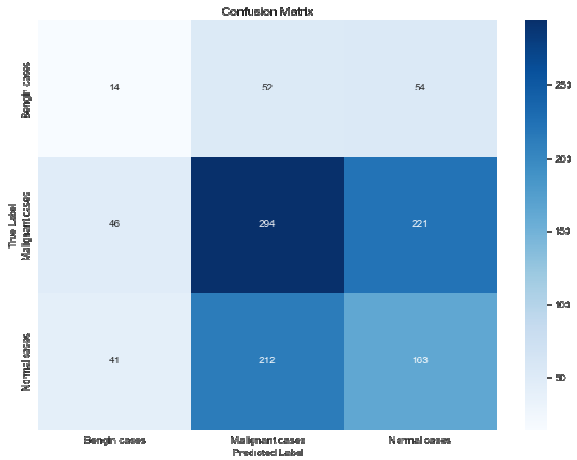

Figure 7 shows the confusion matrices of the dataset IQ-OTH/NCCD

dataset classes. Figure 5 shows the receiver operating characteristic ROC-AUC curve. The X-axis represented a false positive rate and the Y-axis represented a true positive rate. The ROC for all the three classes has been plotted in the figure 5. The CMC curves have been also plotted in the fig.6. The CMC and AUC curves correspond to three classes of IQ-OTH/NCCD datasets for good visualization of the proposed model and performance evaluation. The AUC

curve represents the X-axis as the false positive rate, and the Y-axis represents the true positive rate in fig. 6.

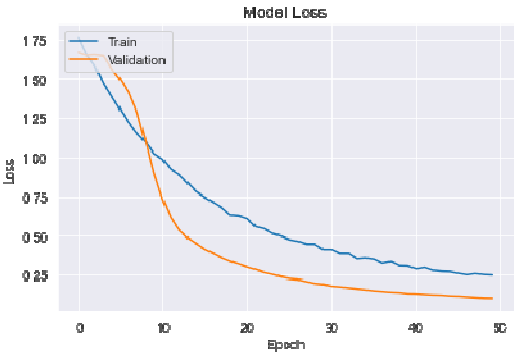

Figure 8 shows the loss curve of the dataset LC25000

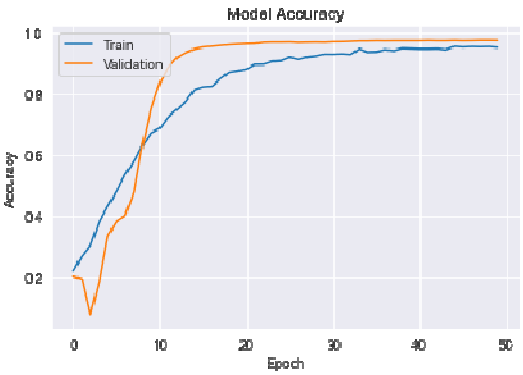

Figure 9. shows the accuracy curve of the dataset LC25000

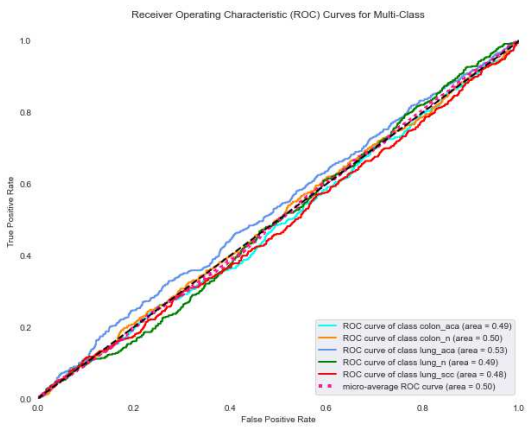

Figure 10 Shows the ROC-AUC curve of the dataset IQ-OTH/NCCD

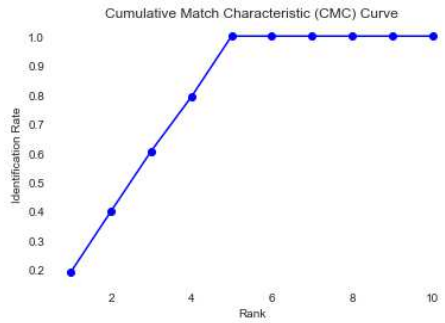

Figure 11 Shows the CMC curve of LC25000 dataset

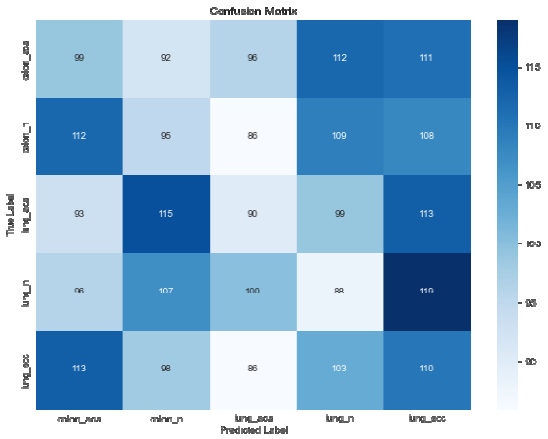

Figure 12 shows the confusion matrices of the dataset IQ-OTH/NCCD

Figures 8 and 9 of the loss and accuracy curve of the LC25000 dataset are given above figure. After plotting the accuracy and loss curve then, calculate the confusion matrices in Figure 12 given below, which represent the x-axis indicating the predicted label and the Y-axis representing the true labels within five classes of LC25000 datasets. Figure 10 shows the receiver operating characteristic ROC-AUC curve. The X-axis represented a false positive rate, and the Y-axis represented a true positive rate, Figure 11 plots the CMC curves. The CMC and AUC curves correspond to five classes of LC25000 datasets for good visualization of the proposed model and evaluating performance. The AUC curve represents the X-axis as the false positive rate, and the Y-axis represents the true positive rate in fig 10 above. The CMC curve X-axis represents the rank Y-axis and presented predicted labels in fig. 11 above figures.

To demonstrate how well our proposed model works for classifying lung cancer, we've compared its performance to other methods in three different tables: Table 6 and Table 7. Table 6 shows the LC25000 dataset, and Table 7 shows the IQ-OTH/NCCD dataset.

Table 6 Performance of the model with classification report

Class ID	*Precision*	*Recall*	*F1-score*	*Accuracy*
colon_aca	0.90	0.90	0.80	99.09%
colon_n	0.99	0.80	0.90	99.09%
lung_aca	0.87	0.98	0.99	99.09%
lung_n	0.90	0.95	0.80	99.09%
lung_scc	0.95	0.87	0.93	99.09%

Table 7 Performance of the model with classification report

Class ID	*Precision*	*Recall*	*F1-score*	*Accuracy*
Bengin class	0.85	0.90	0.87	98.80%
Malignant class	0.80	0.75	0.77	98.80%
Normal class	0.90	0.85	0.87	98.80%

IV. CONCLUSION

Developing and validating a Vision Transformer-based automated model for accurate lung cancer classification marks a significant advancement in histopathological image analysis and diagnostic tools. Our proposed work

demonstrates that Vision Transformer offers superior performance in accurately classifying lung cancer compared to traditional convolutional neural network approaches. The model is robust and validated through extensive testing on to different datasets, ensuring their reliability and generalization across different patient populations and imaging conditions. The highest accuracy achieved in our proposed model is 98.80% for the LC25000 dataset and 99.09% for the IQ-OTH/NCCD dataset with vision transformer based model. The model performance is effective and very efficient with respect to existing models.

References

[1]. J Ferlay, M Ervik, LamF,Colombet M,et. al. Global cancer observatory: cancer today. International Agency for Research on Cancer,Lyon, 2021

[2]. https://www.iarc.who.int/wpcontent/uploads/2018/09/pr263_E.pdf

[3]. WHO Cancer https://www.who.int/news-room/fact-sheets/detail/cancer.Accessed June 2021

[4]. JR Molina, P Yang, SD Cassivi, SE Schild, A AAdjei, "Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship." *Mayo Clinic Proc* Vol. 83, no. 5, pp. 584–594. https://doi.org/10.4065/83.5.584

[5]. S A El-Regaily, SalemMA, Abdel Aziz MH, Roushdy MI, Survey of computer aided detection systems for lung cancer in computed tomography, Medical Imaging Vol. 14, no. 1, pp.:3–18, 2018

[6]. American Cancer Society (2019) Lung Cancer Causes [Online].https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/what-causes.html

[7]. G A Silvestri, M K Gould, ML Margolis, L T Tanoue, D. McCrory, E Toloza, F Detterbeck "Non-invasive staging of non-smallcell lung cancer": *ACCP evidenced-based clinical practice guide-lines. Chest* vol. 132 no. 3, pp-:178S-201S, 2007.

[8]. WD Travis, E Brambilla, M Noguchi, A G Nicholson, K R Geisinger, Y Yatabe, D Yankelewitz, "*IASLC/ATS/ERS international multidisciplinary classification of lung adenocarcinoma*". J ThoracOncol vol. 6 no. 2, pp. :244–285, 2011

[9]. L G Collins, C Haines, R Perkel, R E Enck "*Lung cancer: diagnosis and management*" .AmFam Physician, Vol. 75 no.1, pp.:56–63, 2007

[10]. K H Yu, C Zhang, G J Berry, R B Altman, Ré C, D L Rubin, M Snyder "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. NatCommun vol. 7:12474, 2016

[11]. R L Siegel, K D Miller, A Jemal Cancer statistics, C.A. Cancer J Clin vol. 69 no. 1, pp-734. 2019 https://doi.org/10.3322/caac.21451

[12]. O. Singh, Koushlendra Kumar Singh. "*An approach to classify lung and colon cancer of histopathology images using deep feature extraction and an ensemble method*." International Journal of Information Technology vol. 15, no. 8 , pp-: 4149-4160, 2023.

[13]. R. Mehta, , et al. "*A vision transformer-based automated human identification using ear biometrics*." Journal of Information Security and Applications vol. 78 pp- 103599, 2023:

[14]. R. Sun, , P. Yuexin, and Wenfa Li. "*Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on an Improved Swin Transformer.*" Electronics vol.12., no. 4 pp-10-24, 2023:.

[15]. Fu-Ming Guo, F. Yingfang. "*Zero-Shot and Few-Shot Learning for Lung Cancer Multi-Label Classification using Vision Transformer.*" arXiv preprint arXiv:2205.15290 (2022).

[16]. Z. Jinhua, et al. "*Hepatocellular carcinoma histopathological images grading with a novel attention-sharing hybrid network based on multi-feature fusion*." Biomedical Signal Processing and Control Vol. 86 pp-105-126,2023:.

[17]. Li, Han, et al. "*A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis*." Computers in biology and medicine vol. 151: pp- 106-265, 2022.

[18]. I. Chhillar, , Ajmer Singh. "*A feature engineering-based machine learning technique to detect and classify lung and colon cancer from histopathological images*." Medical & Biological Engineering & Computing, pp- 1-12, 2023:.

[19]. Chen, Chuheng, et al. "*Identifying primary tumor site of origin for liver metastases via a combination of handcrafted and deep learning features*." The Journal of Pathology: Clinical Research (2024).

[20]. T. Saikia, , et al. "*An automatic lung nodule classification system based on hybrid transfer learning approach.*" SN Computer Science Vol. 3.no. 4, 2022:.

[21]. T. Saikia, , et al. "*Classification of lung nodules based on transfer learning with K-Nearest Neighbor (KNN).*" 2022 IEEE international conference on imaging systems and techniques (IST). IEEE, 2022.

[22]. A Esteva, B Kuprel, R.A Novoa, Ko J, Swetter SM, Blau HM, Thrun S "*Dermatologist-level classification of skin cancer with deep neural networks*" .Nature Vol. 542, 7639:pp-115–118, 2017

[23]. R Yamashita, M Nishio, Do RKG, K Togashi "*Convolutional neural networks: an overview and application in radiology*". Insights Imaging Vol. 9,4, :pp.611–629, 2018

[24]. Cancer [online].https://www.who.int/news-room/fact-sheets/detail/cancer , 2020.

[25]. A L Fogel, J C Kvedar *Artificial intelligence powers digital medicine.* N P J Digit Med vol. 1, pp.1-5, 2018

[26]. Gabralla, LubnaAbdelkareim, et al. "*Automated Diagnosis for Colon Cancer Diseases Using Stacking Transformer Models and Explainable Artificial Intelligence*." Diagnostics, vol.13.18: pp.29-39, 2023.

[27]. Chen, Chuheng, et al. *"Identifying primary tumor site of origin for liver metastases via a combination of handcrafted and deep learning features."* The Journal of Pathology: Clinical Research ,2024.

[28]. Chhillar, Indu, and Ajmer Singh. "*A feature engineering-based machine learning technique to detect and classify lung and colon cancer from histopathological images*." Medical & Biological Engineering & Computing pp- 1-12, 2023.