International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Effective Method for Lung Cancer Classification Using Convolutional Neural Network

P. Deepa[1], M. Arulselvi[2], M. Meenakshi Sundaram[3]

Submitted:27/04/2023 **Revised:**28/06/2023 **Accepted:**06/07/2023

Abstract: The incidence of lung cancer has been increasing exponentially in recent years due to hazardous consumption habits and environmental factors. While there are ongoing comprehensive research efforts in the field, the accuracy and efficiency of lung cancer detection remain a challenge. To address this, this study proposes a multi-view aspect model using digital image processing techniques for lung cancer research. The model utilizes Convolutional Neural Networks (CNN) to categorize different types of lung cancer, leveraging the power of image classification capabilities. By employing CNNs, the model aims to enhance the diagnostic accuracy in lung cancer detection. To evaluate the model's performance, several metrics are used, including Matthew's correlation coefficient, Cohen's Kappa score, and log loss. Matthew's correlation coefficient measures the correlation between predicted and actual classifications, providing insights into the overall performance of the model. Cohen's Kappa score assesses the agreement beyond chance between predicted and actual classifications. The log loss metric measures the accuracy of the model's probability estimates. By incorporating these evaluation metrics, this research aims to provide a comprehensive assessment of the proposed multi-view aspect model for lung cancer diagnosis. The goal is to improve the accuracy and efficiency of lung cancer detection, enabling earlier interventions and better patient outcomes.

Keywords: *Convolution Neural Networks, Computer-Aided Diagnosis, Lung cancer Image Dataset Consortium (LIDC)*

1. Introduction

Lung cancer is a deadly disease in underdeveloped countries with a higher mortality rate of 19.4%. Also, it is a dangerous, lethal cancer with the lowest survival rate after prediction and an annual increase in the death rate [1]. To ensure prompt treatment and a higher survival probability, it is critical to discover metastatic tumors at an early stage. On the other hand, the diagnosis procedure is time-consuming and costly because it necessitates human intelligence to make key decisions. It is vital to classify and detect lung cancer early to avoid deaths and boost survival rates. Lung nodules are tiny masses of tissue that might be benign or malignant. Early detection of a tumor will almost always result in the survival of huge people [2].

Convolutional Neural Networks recently attained significant outcomes in a range of appliances together with Lung cancer detection in Deep Learning. Deep learning (DL) algorithms are effectively performed in this lung cancer research [3].

Neural network is acting for identification of cancer cells with usual tissues, which gives a proficient utensil used for structuring an assistive AI based cancer identification. The cancer healing would be an efficient only when the tumor cells are precisely alienated since the usual cells. Categorization of tumor cells and working out of neural network forms the foundation to deep learning support cancer diagnosis. In this methodology, Convolutional Neural Network (CNN) is used to categorize lung carcinoma into any of the following categories adenocarcinoma, huge cell carcinoma, squamous carcinoma, and typical. This CNN model is evaluated with different optimization algorithm, and the presentation of the model is quantified based on the following multiclassification performance metrics such as Matthew's correlation co-efficient, Cohen's kappa score, and log loss.

2. Related Work

Qing Zeng et al [4]. used three deep network models (e.g., DNN, CNN and SAE). These networks are provided for CT image classification with modest modifications for malignant and benign lung nodules. The LIDC-IDRI dataset is to make use of to review these networks.

Kumar et al [5], on hand a method utilizing the Stacked Auto Encoder (SAE), a deep learning technique.Sarfaraz Hussein et al [6] used supervised / unsupervised learning algorithms used algorithms on two diverse tumour

[1]*Research Scholar, Department of CSE Annamalai University, Chidambaram, Email: deepu.prithiv@gmail.com*
[2]*Associate Professor, Department of CSE, Annamalai University Chidambaram, Email: marulcse.au@gmail.com*
Principal, Mahalakshmi College of Engineering, Trichy, Email: bosemeena@gmail.com

diagnosis challenges: pancreatic and lung with 1018 CT and 171 MRI scans and achieves best sensitivity and specificity scores in both issues. And they were also seeking a solution to the baseline question of whether "deep characteristics" are helpful for the categorization of lung tumours without supervision.

Baskar et al. [7] Delta Radiomics uses machine learning techniques to extract the characteristics of cancer nodules. SVM is utilized to measure the lung cancer nodules malignancy that needs to be predicted. The SVM can analyze compact factors in the lungs. The image of a cancer nodule and its categorization was admirably distinguished among the various nodules. As a outcome, SVM is suggested as the finest option, and it has a suitable method for identifying and diagnosing lung cancer.

TafadzwaL. Chaunzwa et al.[8] proposed anon invasive standard CT scan, a radionics technique for predicting the big cell lung cancer tumour histology. A dataset composed of 311 early stage NSCLC patients receives surgical treatment at MGH. Moreover, trained and validated models include two frequent histological types: SCC and ADC. Through an AUC of 0.71 ($p = 0.018$), the CNNs is used to identify tumour histology. The author also discovered that utilizing machine learning classifiers like k-NN was beneficial and AUC sofupto0.71 ($p = 0.017$). SVM based quantitative radionics features produced equal discriminative performance.

Swati et al [9] provides a lung cancer detection framework using AI concepts relying on supervised learning to yield upper precision, particularly when taking up deep learning mechanism. CNN classification is a lung tumour categorization approach. Data acquisition, preprocessing, enhancement, segmentation, feature extraction, neural framework detection are some techniques integrated in the structure. To lay it one more way, a machine learning approach offers an once-in-a-lifetime opportunity to improve lung cancer treatment decision support at a low cost.

Diego et al.[10] suggest computer-aided diagnosis (CAD) systems. DL approaches give impressive outcomes by outperforming the traditional methods. Investigators are currently experimenting with diverse DL approaches to enhance the effectiveness of CAD. Here, the cutting-edge DL algorithms are presented. They are grouped into two types: (1) nodule identification methods to predict candidate nodules from the CT scan, and (2) a comparison of the various methodologies is provided and analyzed.

Wentao Zhu et al. [11] propose using 3D- Multipath VGG Network-based Lung patient computer Tomography (CT) scan pictures to discover and categorize lung modules, as well as to conclude the stage of malignancy of those nodules. U-Net architecture is used to segment CT scan pictures.

P. Khatun and H. K. Rana et al. [12] used ML approaches in healthcare which could be highly beneficial in healing millions of people's illnesses. Researchers have put in a lot of work to discover and deliver early-stage cancer diagnosis information. Various algorithms SVM, k-NN, DT and RF are used in machine learning research to compute the existence of cancer concerning the symptoms displayed by patients. The target is to examine the cause of the three age groups: working class, youth and elderly.

AmjadRehman et al. [13] classify lung malignancies like large cell carcinoma, adenocarcinoma and squamous cell carcinoma. The CT scan is analyzed for texture feature learning using ML approaches like SVM and k-NN. During feature extraction, fusion with patch basis of LBP and DCTis performed.

Mohammed, S. H. M., & Çinar, A [14] lung nodule categorization has been presented the data has utilized of CT image is SPIE AAPM-Lung. It is executed pre-trained Convolutional Neural Networks contain: AlexNet, ResNet18, GoogleNet, and ResNet50 models. The assessment of models is intended by a few matrices are confusion matrix, precision, recall, specificity, and f1-score.

Amjad Khana, Zahid Ansarib [15],[16] Deep Convolutional Neural Network for lung cancer classification by CT images based Lung cancer Image Dataset Consortium (LIDC) for identifying cancerous and non-cancerous lung nodules for appraising precision of categorization enhanced than breathing methods.

3. Proposed Work

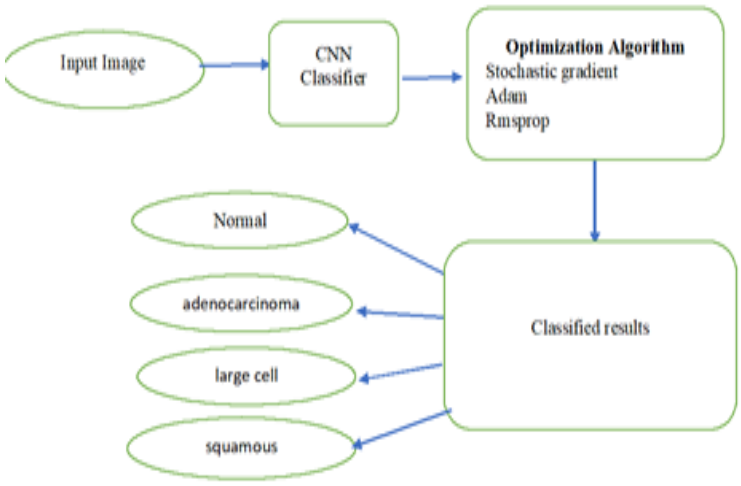


Fig.1 Block diagram of the anticipated work

Figure 1.demonstrates the building block diagram of the projected work. The CNN classifier with the various optimizer algorithms are Root Mean Squared Propagation (RMSProp), Adaptive Moment Estimation (ADAM), Stochastic Gradient, is used to categorize the lung cancer in to one of the four types as Normal, Adeno Carcinoma, Large cell Carcinoma and Squamous Carcinoma

3.1 Dataset

To assess the performance of the classification model, a popularly available lung CT images in the kaggle competition portal is used which consists of 1276 images to train and evaluate the deep neural classifier.

Before training a CNN, data augmentation was used to expand the training set's sample size. First, a 12-degree gap between rotations of 0 and 360 degrees and vertical flipping were added to the dataset. Elastic deformation was also used to enhance images. Where 1276 original images, 3600 augmented images by rotation & flipping, and 600 images generated by deformation, in total 5476 images is used for training. The train test split ratio was fixed at 0.75, thus 319 images is used for testing.

3.2 Network Architecture

Figure 2.shows the CNN network architecture of this approach. A CNN has key layer, hidden layers, and an output layer. However, a hidden layer is comprised with Convolutional layer, a Pooling Layer & completely attached layer. The high-level features are extracted at convolution layer. The input image of dimension 200x400 considered as 2D vector is passed to the first convolution layer of size 100×200@64 after applying batch normalization and in the second convolution layer for determining feature vectors the size of the image is 49x99@64. At the convolutional layer 3, the image is reduced to a dimension of 23x48@64.The pooling layer employs max pooling with stride 2×2 and generates the feature maps through the minimized dimensions. In the dense layer, 1024 neurons are created in the completely connected layer so that the 2D feature map is converted into a 1D feature map wherein the inputs can be mapped with the output by a onto function.

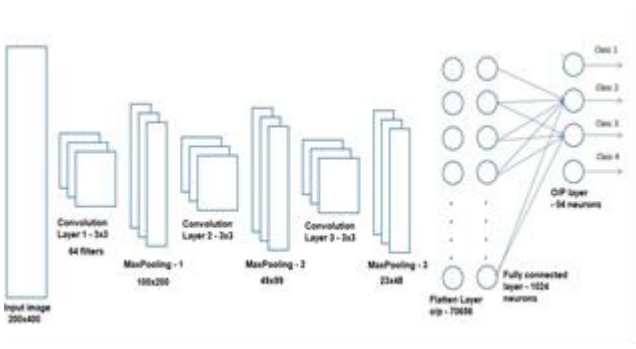


Fig.2 Architecture of CNN

3.3 Optimizer Algorithm

In this work, the CNN classifier with the various optimizer algorithms like the Root Mean Squared Propagation (RMSprop), Adaptive Moment Estimation (Adam), Stochastic Gradient Descent (SGD). All optimizers aim to arrive at the global minima where the cost function has the smallest value. Neural network training starts with random weight values and we get closer to the ideal value every time the gradient is estimated and alter the weights' and biases' values. The cost or error function value would be large before we started training neural network. The cost drops and becomes nearer to the global minimum value with each round of neural network training (identifying gradients and changing the weights and biases). The cost

function is not always as smooth and cost functions are frequently non-convex. Non-convex functions have the drawback that the loss may not ever converge to the global minimum value and that the error function value may become stuck in a local minima.

Gradient descent is an optimization method places a least of an intent function by subsequent the negative gradient of the function. In addition of momentum, the gradient descent optimization procedure with stand the fluctuations of noisy gradients and ride across areas of the seek space are flat. The above permits look for to build up inertia in a particular path in the search space. The equation below expresses the gradient descent optimization with momentum

$$v_{dw} = \beta . v_{dw} + (1 - \beta). dw$$

$$v_{db} = \beta . v_{db} + (1 - \beta). db$$

$$W = W - \alpha . v_{dw}$$

$$b = b - \alpha . v_{dw}$$

Where α - learning rate β −momentum factor whose value is fixed as 0.9. ***W***-Weight, ***b*** -bias value, V***dw*** and V***db*** are the values of derivatives of the gradient.

The Root Mean Squared Propagation (RMSprop) optimizer can be expressed mathematically as below;

$$v_{dw} = \beta . v_{dw} + (1 - \beta). dw^2$$

$$\mathbf{v_{db} = \beta . v_{db} + (1 - \beta). db^2}$$

$$\mathbf{W = W - \alpha . \frac{dw}{\sqrt{v_{dw}} + \epsilon}}$$

$$\mathbf{b = b - \alpha . \frac{db}{\sqrt{v_{db}} + \epsilon}}$$

The V_{dw} may occasionally be extremely close to zero. The value of the model weights might then explode. The denominator has a parameter i.e., epsilon is set to a small value in order to keep the gradients from exploding.

Adaptive Moment Estimation (Adam) optimizer entails a blend of two gradient descent methods, this algorithm is for accelerating the gradient descent algorithm by getting into a concern the exponentially weighted average' of the gradients. Using averages creates the algorithm converge in the direction of the minima in a faster pace.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g2_t$$

Where, m and v are moving averages, g is a gradient and betas are novel pioneer hyper-parameters of the algorithm. The fine default values of 0.9 and 0.999.

3.4 PERFORMANCE METRICS

The performance of developed CNN trained to perform multi classification task was analyzed based on the following performance metrics

Confusion Matrix: A Confusion matrix is for finding the recital of the classifier model by using the table 1.

Accuracy

The accuracy illustrates the degree to which the model precisely identified both positive and negative cases.

Accuracy

$$= \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Classes	Normal	Adeno	Squamous	Largecell
Normal	*True Norma (TP)*	*False Adeno (FP)*	*False Squamous(TN)*	*False Largecell (TN)*
Adeno	*False Normal (FN)*	*True Adeno(TP)*	*False Squamous (FN)*	*False Largecell (FN)*
Squamous	*False Normal (TN)*	*False Adeno(FP)*	*True Squamous(TP)*	*False Largecell (TN)*
Largecell	*False Normal (TN)*	*False Adeno(FP)*	*False Squamous (TN)*	*True Largecell(TP)*

Table 1: Confusion Matrix of Multiclass Classification

Matthew's Correlation Co-efficient (MCC): With reference to the confusion matrix used for evaluating the binary classifiers performance, the MCC is an optimal metric suitable for evaluating the performance of a multiclassification model irrespective of the class imbalance. The coefficient value ranges between 0 and 1, a value closer to one indicates a good classification model and a value closer to zero indicates poor training result.

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

Log Loss Function: It is an error function that estimated the log loss termed as cross-entropy loss and hence the classifier that yields minimum value of the log function is the optimal one. It estimates the loss value based on the probability scores. Also, it can be inferred that it calculates probability membership scores. This takes the uncertainty of the model in to account. The metrics which uses prediction labels for quantifying the model performance certainly hides the uncertainty of the model. It highly penalizes the test instances for which the model predicted class membership has lower scores.

Where,M – number of classes present in the dataset, Y– a binary value indicating the correct classification for observation'o', P – predicted probability score indicating observation o is of type c.

Cohen's Kappa Co-efficient: For qualitative (categorical) items, Cohen's kappa coefficient (k) is a statistic for evaluating inter-rater consistency as well as intra-rater consistency. Since it judge the prospective that the accord could have occurred by possibility, it is usually alleged to be an additional consistent dimension than a simple % conformity estimate.

$$k = \frac{p_o - p_e}{1 - p_e}$$

The relative concordance between actual and expected values, denoted by p_o, was observed. This represents the sum of any confusion matrix's diagonal cells divided by the sum of its non-diagonal cells. p_e is the likelihood that true values and false values will coincide by accident.

4 Experimental Results And Discussion

The given 200x400 was the input size for our CNN architectures. The batch 1 of Convolutional layer, the input size is 200*400*64, Max pooling layer size is 100*200*64 and output parameters are 1792. The batch 2 of Convolutional layer, the input size is 98*198*64, Max pooling layer size is 49*99*64 and output parameters are 36928. The batch 3 of Convolutional layer, the input size is 47*97*64, Max pooling layer size is 23*48*64 and output parameters are 36928. The dense parameters are 72351744. The utilization of total parameters are 72,436,356 and trained parameters are 72,433,924. The images were resized using a bi-cubic method. The CNN architectures were exposed to 50 epochs of training. The image samples present in the training set was divided into two separate, non-overlapping subsets for training and validation. Flip and rotate image augmentations as well as elastic deformation was performed to the training images of the dataset in order to produce more training images for the CNN training. The same training and validation sets were used to train the CNN architecture with various initial random weight values. 0.0001 was the initial learning rate, and training period batch size was fixed as 16. The The final classification layer used a softmax function for activation because there are four classes of image samples. Accuracy was determined for performance evaluation using forecasts on independent, hidden test data.

Classes	Normal	Adeno	Squamous	Largecell
Normal	54	4	0	10
Adeno	0	110	0	0
Squamous	0	0	90	0
Largecell	0	0	0	51

Table 2: Confusion Matrix of Multiclass Classification

	SGD	RMSProp	ADAM
No augmentation	78.68%	78.05%	76.64%
Rotation and Flipping (a)	84.98%	83.54%	81.86%
Elastic Deformation (b)	82.56%	81.49%	80.98%
Both (a) and (b)	88.30%	87.26%	85.65%

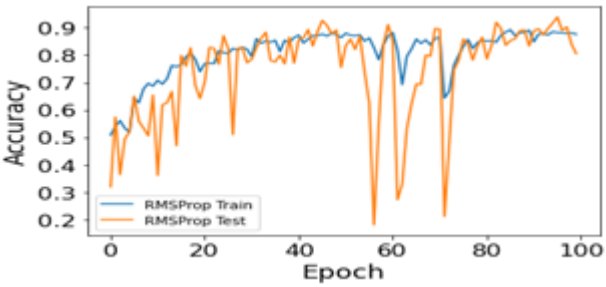
Table 3: Performance accuracy of CNN under various optimizers

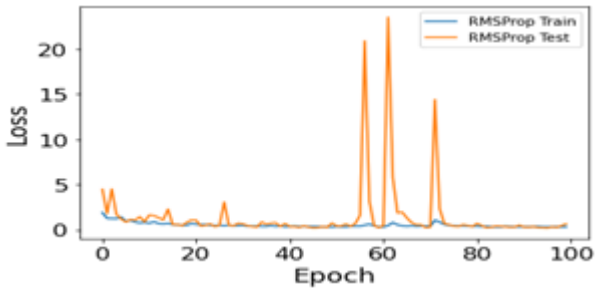
	Matthew's Correlation Co-efficient	Cohen's Kappa Co-efficient	Log Loss
SGD	0.9406	0.9385	0.1632
RMSProp	0.9242	0.9168	0.1916
ADAM	0.9036	0.8954	0.2108

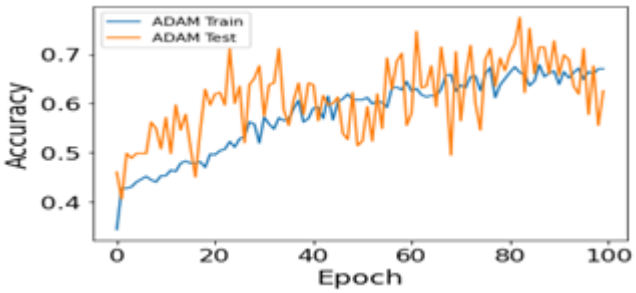
Table.4. Summary of Performance Metrics

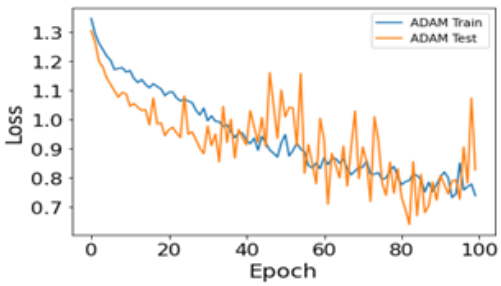
	Normal	Adeno	Squamous	Largecell
Training	1680	1345	1298	1153
Testing	54	114	90	61

Table.5. Classifier result for Multiclass classification

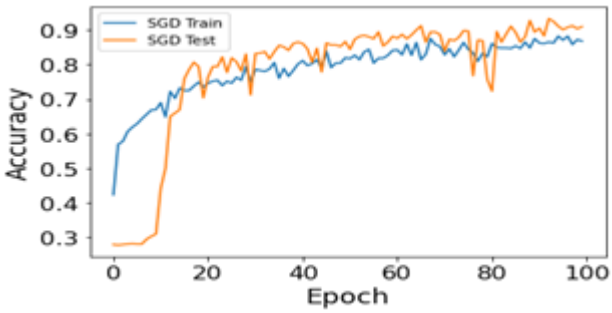


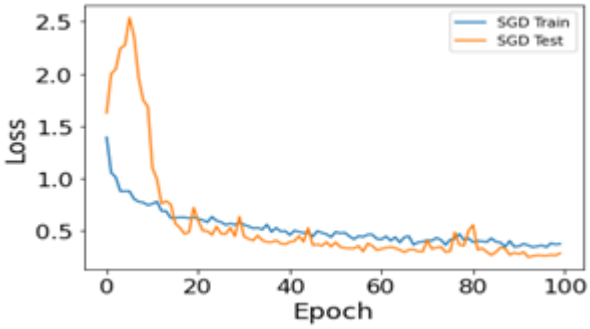






(a) RMSProp **(b) ADAM**





(c) SGD

Fig 3. Performance Comparison of CNN under various optimizers

5 Conclusion

In conclusion, it was observed that flipping and rotating images was a very effective method of augmentation of images. Deep neural network classification of lung images using SGD is a convincing method that greatly improves classification outcomes. Instead of 3D volumes, applying 2D slices for our investigation. Although our 2Dmethod has some information deviation compared to 3D.Our outcomes are still significantly superior to those obtained by employing the 3D method described in other literatures. Through training and testing methodology to find SGD that is highly predictive of lung image classification for diagnosis and treatment of lung cancer given the small limitations of this work.

References

[1] Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; Choi, B.,et al., End-to-End Lung Cancer Screening withThree-Dimensional Deep Learning on Low-DoseChest Computed Tomography. Nat. Med. 2019,25, 954–961, doi:10.1038/s41591-019-0447-x.

[2] Rehman, M. Kashif, et al., "Lung Cancer Detectionand Classification from Chest CT Scans usingMachine Learning Techniques," 2021 Artif. Intell.Data Anal. CAIDA 2021, pp. 101–104,2021,<doi:10.1109/CAIDA51941.2021.9425269>

[3] T.L.Chaunzwa et al., "Deep learningclassificationof

lung cancer histology byCT images," Sci. Rep.,vol. 11, no. 1, pp. 1–12, 2021, <doi:10.1038/s41598-021-84630-x>.

[4] QingZeng, "Using Deep Learning forclassificationof Lung Nodules on computed TomographyImages, 2017.

[5] D. Kumar, A. Wong, and D. A. Clausi, "Lung noduleclassification using deep features in CT images,"in 12th Conferenceon Computer and Robot Vision(CRV), pp.133–138, IEEE, 2015.

[6] Sarfaraz Hussein et al, "Lung and pancreatic Tumorcharacterization in the Deep learning Era: NovelSupervised and Unsupervisedlearning approaches,"IEEE Trans. Med.Imaging, vol. 38 (8), pp. 1777–1787, 2019.

[7] S. Baskar, "Classification System for Lung CancerNodule Using Machine Learning Technique andCT Images," Proc. 4th Int. Conf. Common. Electron.Syst. ICCES 2019,pp. 1957–62, 2019.

[8] TafadzwaL. Chaunzwa, "Deep learning classificationof lung cancer histology using CT images," Sci. Rep.,vol. 11(1), pp. 1–12, 2021.

[9] Swati Mukherjee and S.U.Bohra "Lung cancerdisease diagnosis using machine learning approach,"Proc. 3rd Int. Conf. Intell. Sustain. Syst. ICISS 2020,pp. 207–211, 2020.

[10] Riquelme, "Deep Learning for Lung Cancer NodulesDetection and Classification in CT Scans," vol.1(1), pp. 28–67, 2020.

[11] Zhu, W.; Liu, C, et al., Deep Lung: Deep 3D Dual-Path Nets for Automated Pulmonary NoduleDetection and Classification ar Xiv:1801.09555,2018.

[12] M. H. Jony, F. TujJohora, P. Khatun and H. K. Rana,"Detection of Lung Cancer from CT Scan Imagesusing GLCM and SVM", 2019 1stInternationalConference on Advances in Science Engineering and Robotics Technology (ICASERT), pp. 1-6, 2019

[13] Amjad Rehman, "Lung Cancer Detection and Classification from Chest CT Scans using MachineLearning Techniques," 1st Int. Conf. Artif. Intell.Data Anal. CAIDA, pp. 101-104, 2021.

[14] Mohammed, S. H. M., & Çinar, A., "Lung cancerclassification with Convolutional Neural NetworkArchitectures", Qubahan Academic Journal, 1(1),33–39.(2021). https://doi.org/10.48161/qaj.v1n1a33

[15] Amjad Khana, Zahid Ansarib, "Identification ofLung Cancer Using Convolutional Neural Networksbased Classification", Turkish Journal of Computerand Mathematics Education Vol.12 No.10 (2021),192-203

[16] Chitra, T., Sundar, C., & Gopalakrishnan, S. (2022). Investigation and Classification of Chronic Wound Tissue images Using Random Forest Algorithm (RF). International Journal of Nonlinear Analysis and Applications, 13(1), 643-651. doi: 10.22075/ijnaa.2021.24438.2744

[17] Martinez, M., Davies, C., Garcia, J., Castro, J., & Martinez, J. Machine Learning-Enabled Quality Control in Engineering Manufacturing. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/122

[18] Anupong, W., Azhagumurugan, R., Sahay, K. B., Dhabliya, D., Kumar, R., & Vijendra Babu, D.(2022). Towards a high precision in AMI-based smart meters and new technologies in the smart grid. Sustainable Computing: Informatics and Systems, 35 doi:10.1016/j.suscom.2022.100690