Lung Cancer Classification : A New EnhancedLearning Capability of CNN

V Likitha
Department of CSE, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, Guntur-522502, India
K saikumar, ECE, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram 522302,

Andhra Pradesh, India

ABSTRACT:

Cancer is the uncontrollable cell division of abnormal cells inside the human body, which can spread to other body organs. Cancer survival rate of lung cancer is only 19%. There are various methods for the diagnosis of lung cancer, such as X-ray, CT scan, PET-CT scan, bronchoscopy andbiopsy. However, to know the subtype of lung cancer based on the tissue type H and E staining is widely used, where thestaining is done on the tissue aspirated from a biopsy. Studies have reported that the type of histology is associated with prognosis and treatment in lung cancer. Therefore, early and accurate detection of lung cancer histology is an urgent need and as its treatment is dependent on the type of histology, molecular profile and stage of the disease, it is most essential to analyse the histopathology images of lung cancer. Hence, to speed up the vital process of diagnosis of lung cancer and reduce the burden on pathologists, Deep learning techniques are used. These techniques have shown improved efficacy in the analysis of histopathology slides of cancer. Several studies reported the importance of convolution neural networks (CNN) in the classification of histopathological pictures of various cancer types such as brain, skin, breast, lung, colorectal cancer. In this study tri-category classification of lung cancer images (normal, adenocarcinoma and squamous cell carcinoma) are carried out by using ResNet 50, VGG-19, Inception_ResNet_V2 and DenseNet for the feature extraction and triplet loss to guide the CNN such that it increases inter-cluster distance and reduces intra-cluster distance.

INTRODUCTION:

There are various methods for the diagnosis of lung cancer, such as X-ray, CT scan, PET-CT scan, bronchoscopy and biopsy. However, to know the subtype of lung cancer based on the tissue type H and E staining is widely used, where thestaining is done on the tissue aspirated from a biopsy [1]. Hematoxylin (H) has a deep purple colour, stains nucleic acids in the cells and Eosin (E) have pink colour, and it stains proteins. (Fischer et al, 2008). Studies have reported that the type of histology is associated with prognosis and treatment in lung cancer (Hirsch et al, 2008; Itaya et al, 2007; Weiss et al, 2007) [2]. Recent advances in genomic studies paved the path to personalized medicine for lung cancer patients (Travis et al, 2021; Galli and Rossi, 2020). Therefore, early and accurate detection of lung cancer histology is an urgent need and as its treatment is dependent on the type of histology, molecular profile and stage of the disease, it is most essential to analyse the histopathology images of lung cancer[3]. However, manual analysis of histopathology reports is time-consuming and subjective. With the advent of personalized medicine, pathologists are finding it difficult to manage the workload of dealing with a histopathologic cancer diagnosis. Hence, to speed up the vital process of diagnosis of lung cancer and reduce the burden on pathologists [4], Deep learning techniques (Baranwal et al, 2019, Tripathi et al, 2013, Kumud et al., 2015 and singh et al, 2020) are used. These techniques have shown improved efficacy in the analysis of histopathology slides of cancer (Litjens et al, 2016) [5].

Cancer is the uncontrollable cell division of abnormal cells inside the human body, which can spread to other body organs. The process of transformation of normal cells into cancerous cells due to genetic alteration is known as Carcinogenesis as shown in Figure 1 [6]. The process of carcinogenesis occurs in three phases. The first is the Initiation phase, where any alterations that occur in the normal cell due to gene mutation can cause a change in gene expression and even deletion of a part of Deoxyribonucleic acid (DNA) sometimes [7]. If these changes skip the repair mechanism during the cell cycle, then the cell with altered genes remains as it is. In the Promotion phase, which is the second phase, the altered cell starts proliferation [8]. In the final stage, the Progressive phase the cells start proliferating aggressively by number, size, and form primary tumors. In this stage, the cells become invasive and metastatic. Phases of carcinogenesis is shown in Figure 2 (Chegg.com, 2021

ANALYSIS OF PREVIOUS RESEARCH:

In the next few decades, cancer is expected to be the leading cause of death and is one of the biggest threats to human life (Tang et al, 2009). To improve the efficiency and speed of cancer diagnostics, Computer-aided diagnosis (CAD) was applied to the analysis of clinical data. There has been vast development in the field of CAD and many machine learning techniques are developed for the diagnosis purpose. Among all machine learning techniques, neural networks have shown increased performance in the detection of medical images. In the classification of lung cancer images, different CNN algorithms are used to improve the accuracy of the prediction and classification. Such accurate predictions aid doctors by reducing the workload and prevent human errors in the process of diagnosis [9]. a) Computer aided diagnosis in medicine: Computer-aided diagnosis (CAD) is cuttingedge technology in the field of medicine that interfaces computer science and medicine. CAD systems imitate the skilled human expert to make diagnostic decisions with the help of diagnostic rules. The performance of CAD systems can improve over time and advanced

CAD can infer new knowledge by analysing the clinical data [10]. To learn such capability the system must have a feedback mechanism where the learning happens by successes and failures. During the last century, there is a dramatic improvement in human expertise and examination tools such as X-ray, MRI, CT, and ultrasound. With the discovery and study of new diseases and their progression, the diagnosis has become difficult and more complex. Various factors such as complex medical diagnosis, availability of vast data pertinent to conditions and diseases in the field of medicine, increasing knowledge on diagnostic rules, and the emergence of new areas such as AI, machine learning, and data mining in the field of computer science has led to the development of CAD (Yanase and Triantaphyllou, 2019a). Quantitative analysis of pathology images has gained importance among researchers in the field of pathology and image analysis. There is clearly a need for quantitative image- based evaluation of pathological slides as the diagnosis is based on the opinion of pathologists. CAD can reduce the burden on pathologists by filtering out the benign cancer images so that the pathologists can focus on more complicated images that are difficult to diagnose and suspicious. Quantitative analysis of pathology images not only helps in diagnosis but also in medical research (Gurcan and Boucheron, 2019). At many hospitals in the United States CAD has become a part of routine clinical work for screening mammograms for the detection of breast cancer (Freer and Ulissey, 2001; Doi, 2007). In the fields of radiology and medical imaging, CAD has become the major research subject (Doi, 2007). These are cost-effective and can be used for the early detection of disease. Diseases like cancer are very aggressive when detected at later or advanced stages, hence screening and detection of such disease can avoid unnecessary invasive procedures for the treatment of the disease. Moreover, these models can eliminate human errors such as the detection of microcalcifications and help to improve the workflow of diagnostic screening procedures (Nishikawa et al, 2012; Yanase and Triantaphyllou, 2019a).

PROPOSED RESEARCH WORK:

To classify the lung cancer images, the dataset is obtained from LC25000 Lung and colon histopathological image dataset which is already augmented data having 5000 images in each class of lung cancer image set comprising three classes. This dataset is pre-processed using python tools and features are extracted by CNN techniques, later the model is created and evaluated. Various CNN techniques are used to compare and classify the images. Complete flow of proposed method is shown in Figure 4. <Figure 4 here> a) Dataset description:

Data is drawn from the LC25000 Lung and colon histopathological image dataset, which consists of 5000 images each in three classes of benign (normal cells), adenocarcinoma and squamous carcinoma cells (both are cancerous cells). The dataset is HIPA A compliant and validated (Borkowski et al, 2019). The original images obtained are only 750 images in total and the size of the images are 1024 x 768 pixels, where each category gets 250 each. These images are cropped to 768 x 768 pixels using python and expanded using the augmentor software package. Thus, the expanded dataset contains 5000 images in each category. Augmentation is done by horizontal and vertical flips and by the left and right rotations (Borkowski et al, 2019). The sample images for each category are shown in Figure 5. <Figure 5 here> b) Data Pre- processing: Data pre-processing is an essential step, which helps in improving the quality of the images and it include data preparation, data normalization, data cleaning, and data formatting. Data preparation aids in the transformation of data by modifying it into the appropriate format. Whereas data normalization makes a different image format into a regular format where all the images are uniform while in data transformation, the data is compressed (Zubi and Saad, 2011). As the images are already augmented, ImageDataGenerator which is imported from Keras. Preprocessing, image class used for the preprocessing of the image dataset. A total of 15000 images are used for the train-test split, in which 80% of the images are used for training and 20% for validating the data. c) Feature extraction: Feature extraction is used to decrease the model complexity where important features are recognized from the images. For the knowledge extraction from images, not all the features provide interesting rules for the problem. This is the major step where the model performance and effectiveness are dependent. To extract such features as color, texture, and structure, image- processing techniques are used. This can be achieved by localizing the extraction to small regions and ensuring to capture all areas of the image (Zubi and Saad, 2011). For feature extraction, ResNet 50(He et al, 2016), VGG19(Munir et al, 2019), Inception_ResNet_V2 (Xie et al, 2019; Kensert, Harrison and Spjuth, 2019), DenseNet121(Huang, Liu, Van Der Maaten, and Weinberger, 2017; Chen, Zhao, Liu and Lin, 2021) is used. D) Loss function: For a machine learning model to fit better while training the neural networks, loss function acts as a major key for adjusting the weights of the network. During the back propagation while training, loss function penalizes the model if there is any deviation between the label predicted by model and the actual target label (https://ieeexplore.ieee.org/abstract/document/8943952). Hence the use of loss function is very critical to achieve better model performance. Triplet loss is used as loss function in this study.

RESULT AND ANALYSIS :

All four CNN architecture models have been trained using specific and fine-tuned parameters to achieve better model performance. Initially pre-trained CNN architecture is used to classify the lung cancer cells. In these models' cross entropy is used as loss function. VGG19 model is trained by adding two hidden layers with embeddings 256 and 128 with ReLU as an activation function and for the final output layer softmax is used as activation function. For this model cross entropy is used to calculate the loss over 18 batch size. When the model is trained with 30 epochs with Adam as an optimizer (in default setting), it showed validation loss of 0.196. The performance of the model has shown accuracy of 92.1%, precision of 92.5%, recall of 92.1% and f1 score of 92.04% on validation dataset. Similarly, ResNet50 model is trained using the same number of hidden layers as VGG19. All the parameters are same for both the models and when the model is trained for 30 epochs the validation loss showed by the model is 0.03. Among all ResNet has shown improved performance when compared to VGG19 model. This model showed accuracy, precision, recall and f1 score of 99%. Inception-ResNetv2 is trained using two layers, in which one is global average pooling and the other one is dense layer with 1024 embeddings. The activation layer used for the hidden layer is ReLU and for the output layer is softmax. When the model is trained for 30 epochs with Adam as optimizer in default, the validation loss of the model is 0.008. The performance of this model is much better than other models, where test accuracy, precision, recall and f1score is 99.7%. Lastly, DenseNet121 model which is trained with two hidden layers of 1024 and 500 embeddings with Adam in default setting as optimizer has shown validation loss of 0.01. After evaluation of this model on test data the accuracy, precision, recall and F1score is 99.4%. These evaluation metrics are shown in Table 2 for comparison. All the four CNN architecture Inception-ResNetv2 model has shown improved performance and classified benign tissue images from cancer images without any misclassifications. The only misclassification happened is between the subclasses of lung cancer images as shown in Figure 6. Even validation loss is also very minimum for this model.

CONCLUSION:

CNN models have shown to increase accuracy with fine tuning of hyper parameters. Various CNN architectures are compared in the study to get better accuracy and to compare which architecture gives better performance for this dataset. Model performance of all four CN. Although the pre-trained models are available, fine-tuning of these models are necessary to obtain desired results. In this study Inception-ResNetv2 has shown a very hightest accuracy rate of 99.7% when compared to other models where the accuracy of VGG19, ResNet50 and DenseNet121 are 92,99 and 99.4% respectively. When the triplet neural network model is trained on these four pre-trained models DenseNet121 achieved test accuracy of 99.08% which is the highest of all other four. Test accuracies of other three models are 97.69, 96.2, 97.04% for VGG19, ResNet50 and Inception-ResNetv2 respectively. The obtained model with high accuracy has significantly classified cancerimages from non-cancerous images which is a crucial step in cancer diagnosis. There were no misclassifications among cancer and non-cancer images. Only very few misclassifications happened among the two lung cancer subtypes, that is adenocarcinoma and squamous cell carcinoma. Although the image aspect ratio of image trained triplet neural networks is low, that is 128×128×3 and batch size is 16 due to GPU constraints, the triplet network model has shown better performance.

REFERENCE:

[1] Agarwal, N., Balasubramanian, V. N., & Jawahar, C. V. (2018). Improving multiclass classification by deep networks using DAGSVM and Triplet Loss. Pattern Recognition Letters, 112, 184–190.

[2] Abdel-Zaher AM, Eldeib AM (2016) Breast cancer classification using deep belief networks. Expert Syst Appl 46:139–144.

[3] Agarwal, N., Balasubramanian, V. N., & Jawahar, C. V. (2018). Improving multiclass classification by deep networks using DAGSVM and Triplet Loss. Pattern Recognition Letters, 112, 184–190.

[4] Akkus, Z., Ali, I., Sedlar, J., Agrawal, J. P., Parney, I. F., Giannini, C., & Erickson, B. J. (2017). Predicting deletion ofchromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. Journal of Digital Imaging, 30(4), 469–476.

[5] Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N (2016) Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 35(5):1313–1321

[6] Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). Improving the prediction of survival in cancer patients by using machine learning techniques: Experience of gene expression data: A narrative review.
Iranian Journal of Public Health, 46(2), 165–172

[7] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 686, pp.394–424. Page 12 of 25 Page 12 of 25

[8] Brennan, T.A., 2004. Medical malpractice. The New England Journal of Medicine, 350(3), p.283.

[9] Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Alzheimer's Disease Neuroimaging Initiative. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage, 111, 562–579.

[10] Bukhari, S.U.K., Syed, A., Bokhari, S.K.A., Hussain, S.S., Armaghan, S.U. and

Shah, S.S.H., (2020) The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning. bioRxiv, p.2020.08.15.20175760.