Medical Imaging Process & Technology (2023) Volume 6 Issue 1
doi: 10.24294/mipt.v6i1.2712

ORIGINAL RESEARCH ARTICLE

Identifying the probability of genetic mutations in lung cancer using predictive and prognostic biomarkers from histopathological images

Lokeswari Y. Venkataramana[*], D. Venkata Vara Prasad, G. V. N. Akshay Varma, Chitraju Vishnusree

Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai 603110, India

*** Corresponding author:** Lokeswari Y. Venkataramana, lokeswariyv@ssn.edu.in

ABSTRACT

Background: Lung cancer is the highest deadliest disease and second largest disease being diagnosed worldwide. In the age of precision medicine, determining a patient's genetic status is critical. Finding the percentage of gene mutation of a particular biomarker will help in targeted therapy of a patient at an early stage. **Objective:** Histopathology images are larger in size which needs to be converted into smaller tiles for the computational purpose. Deep Learning Techniques could be applied on this huge number of histopathological images to derive the probability of gene mutation occurrence in predictive and prognostic biomarkers of lung cancer. **Methods:** In this work, a deep learning convolutional neural network (CNN) model (InceptionV3) is trained on histopathology images obtained from The Cancer Genome Atlas (TCGA) to accurately predict the mutated genes in lung adenocarcinoma. The convolutional neural network-based model predicts 10 major genetic mutations percentage, i.e., EGFR, FAT1, FAT4, KEAP1, KRAS, LRP1B, NF1, SETBP1, STK11, TP53. **Results:** InceptionV3 predicted the probability of gene mutation from the histopathology images and categorized the genes as predictive and prognostic. InceptionV3 yielded an accuracy of 82.36% and cross entropy of 37.62%. **Conclusion:** InceptionV3 was trained on histopathology images to predict gene mutations with an accuracy of 82%. Prediction of gene mutations with different CNN models like AlexNet and ResNet can be explored further.

Keywords: lung cancer; biomarker; deep learning; multi label classification and histopathology images

ARTICLE INFO

Received: 11 September 2023
Accepted: 13 September 2023
Available online: 26 December 2023

1. Introduction

Biomarkers are the indicators of biological state. A biomarker could be either blood or urine samples, saliva, tissue from the organ of the body. Biomarker shows the deviations from the actual values of blood or urine samples collected from the patients which serves as an indicator for diagnosing the disease and check how the body reacts to the treatment. It is also known as molecular marker and signature molecule.

There are two major types of biomarkers:

- Predictive Biomarker (biomarkers of exposure), which are employed in risk prediction.
- Prognostic Biomarker (biomarkers of disease), which are used in screening, diagnosis and monitoring of disease progression.

Lung cancer is a type of cancer that occurs in the lungs and spreads to other parts of the body. It is the leading cause of cancer death worldwide. Global lung cancer mortality worldwide is 1,796,144 in the year 2020. Smoking is the main reason for lung cancer and the people who are affected by lung cancer due to smoking is 80% and nonsmokers is 20%. Other reasons for lung cancer are exposure to certain metals and air pollution.

There are two types of lung cancer:

- small cell lung cancer (SCLC)
- non-small cell lung cancer (NSCLC)
    - adenocarcinoma (LUAD)
    - squamous cell carcinoma (SCC)
    - large cell carcinoma (LCC)

There are four stages of lung cancer. They are mentioned as follows:

Stage I: Cancer is in lung tissues but not in lymph nodes.

Stage II: Lymph nodes close to the lungs may have been affected by the cancer.

Stage III: Lymph nodes and the center of chest have been further affected by its spread.

Stage IV: Body is covered in a large amount of cancer. Brain, Bones, or Liver may have been affected.

Note: Adenocarcinoma is most common type and starts in the mucus making gland cells in the lining of the airways.

In the field of genetics, a genetic biomarker (also known as genetic marker) is a DNA sequence that causes disease or is associated with susceptibility to disease.

Most prominent predictive biomarkers are:

- ALK (Anaplastic lymphoma kinase) gene rearrangements and overexpression
- EGFR (Epidermal growth factor receptor) gene mutations
- KRAS (Kirsten RAt Sarcoma virus) gene mutations
- Programmed death ligand 1(PD-L1)
- ROS1(c-ros oncogene) gene rearrangement
- TP53(Tumor protein 53)

Most prominent prognostic biomarkers are:

- BRAF (B-Raf proto-oncogene) v600 mutations
- STK11(Serine/Threonine kinase 11)
- FAT1(FAT atypical cadherin 1)
- SETBP1(SET binding protein 1)

Related work on processing images with machine learning and deep learning techniques were discussed. The challenges involved in extracting features from images and matching them for analysis were discussed in the work[1]. Deep Learning techniques could be applied to analyze thermal images and predict the early-stage oral mucositis. The accuracy of 82% was obtained in predicting the oral mucositis in patients with locally advanced head-and-neck squamous cell carcinoma (HNSCC)[2].

The light radiations scatter in the biological tissues obtained using optical tomography method for early diagnosis of breast cancer. The light propagation in the biological tissues were comparatively studied using two methods namely (i) Stochastic models like Monte Carlo methods and (ii) Deterministic models using Finite element methods[3]. Image processing techniques were applied to process the visual image and convert into an audio for the blind. To recognize the image described by the audio, training process must be done which would aid blind people[4]. Prediction of enhancer in genomic data using machine learning techniques was discussed in the work[5]. To estimate the position and orientation of the camera from the 2D images was carried out in the work[6]. The digital video was divided into number of key frames, the features such as shape, texture and variability in intensity levels were extracted using image processing techniques. The different side effects of radiations of electromagnetic spectrum causes certain types of skin cancer. Thermal imaging should be the most preferred method of diagnosis for humans[7].

The motivation of this research work is to exploit the image processing techniques with Deep learning methods to predict the probability of gene mutations in lung cancer from histopathology images and classify the mutations as predictive or prognostic genes mutations. This study will aid doctors to identify the genes to be concentrated for diagnosis and treatment of lung cancer for each patient.

2. Related work

A survey on relevant work on predicting the gene mutations from gene expression data was carried out which gives knowledge about classifying cancer and prediction of survival.

Amini et al.[8] discusses the combinations for various radiomics modalities towards overall survival prediction in Non-Small Cell Lung Cancer NSCLC patients, single-modality PET and CT and multimodality PET/CT fusion using feature selection and machine learning models. Highest performance was achieved using the gradient boosting linear model. The only limitation is small sample size. An Automatic TPS (Tumour proportion score) assessment algorithm named ATPSS finds tumour proportion score (TPS) for PD-L1 expression from NSCLC[9]. Tumour Region Segmentation and Nuclei Instance Segmentation are applied on whole slide images. The results obtained are Sensitivity of 90.1% and a specificity of 96.9% and a Pearson's Coefficient of 0.91.

Chang and his colleagues' studies[10] predicts the ALK rearrangement status in LUAD using a machine learning model that incorporates PET/CT radiomic data as well as clinical variables. The PET/CT radiomic model and the combined model do not significantly differ from one another.

Chen et al.[11] applied deep-learning techniques on histopathology images to classify images into benign(non-cancerous) and malignant(cancerous) and assessing histopathology grade. The mutation status of 4 of 10 selected genes were predicted for liver cancer. 96.0% accuracy for benign(non-cancerous) and malignant(cancerous) categorization and 89.6% for tumour differentiation that was good, moderate, or poor. Four biomarkers are predicted: CTNNB1, FMN2, P53, and ZFX4. High performance level of model is observed at validation set.

Fu et al.[12] used deep learning techniques to extract histopathological patterns and classification of cancer from normal tissue types. Transfer or weekly supervised learning, InceptionV4, PC-CHiP analysis is used to find associations between histopathological features and genomic driver alterations. Predicted several mutations across 28 cancers. AUC is 0.99, accuracy is high. Gao et al.[13] provided insights on multi-omics and machine learning for improving the prognosis of lung cancer are provided. Korpanty et al.[14] reviewed advances in targeted treatment of LUAD with respect to five prominent biomarkers – EGFR, ALK, MET, ROS-1, and KRAS. Kourou et al.[15] outlined concepts of machine learning that have been proposed in previous years. Combining feature selection and classification for multidimensional data provides insights in cancer research. Mayer et al.[16] ALK and ROS1-fusions are detected from scanned hematoxylin and eosin (HE) whole slide images of NSCLC patients using an advanced convolutional neural network (CNN). Two step training approach – Unsupervised training step and semi-supervised fine-tuning step. Morgado et al.[17], the radiomic features extracted from the CT images were used to predict EGFR mutation. Six ML algorithms were trained, and performance metrics are applied respectively. Murchan et al.[18] discussed the applications of deep learning techniques on histopathology image. In addition, limitations associated with histopathology images are explained. Ni et al.[19] identifies TOP2A, CCNB1, CCNA2, UBE2C, KIF20A, and IL-6 as impending genes associated with the pathogenesis and prognosis of NSCLC. Patil et al.[20] uses machine learning methods in cancer prediction and prognosis. To study cellular pH, a variety of methods have been developed, most of which rely on fluorescence indicators and decorated nanoparticles. Drawback is complex multi-step protocols for nanoparticle synthesis and functionalization. Photo bleaching affects cell physiology and fluorescence imaging methods.

Roman-Canal et al.[21] identified miRNA-1-3p, miRNA-144-5p and miRNA 150-5p as confirming biomarkers for diagnosing the lung cancer. Santarpia et al.[22] gave an overview of challenges in applying deep learning techniques for clinical interpretation from the circulating biomarkers, blood and other biologic fluids. Seijo et al.[23] gave an overview of use of molecular biomarkers in the lung cancer screening setting. Research challenges in the field of biomarkers in the context of lung cancer screening are described. Shiri et al.[24] investigates the effect of harmonization on the effectiveness of CT, PET, and combined PET/CT radiomic characteristics for mutation prediction. Image level fusion, Segmentation, Feature Extraction, Model Evaluation—Univariate and Multivariate Analysis are used. Size of the dataset is limited. Combat harmonization method cannot translate new feature sets using transform methods. In KRAS prediction after combat harmonization, most aspects showed no substantial variation in performance unlike EFGR prediction. Combat Harmonisation has a greater influence on EFGR features than KRAS status prediction, although its effect is feature dependent.

Silva et al.[25] worked on pre-trained encoder for feature extraction and classified EGFR mutations using Multi Layer Perceptron (MLP). Song et al.[26] combined clinical, conventional CT & radiomic features are used to predict the ALK mutations in LUAD patients non-invasively. VOIs Delineation, Radiomic Feature Extraction, DB-SCAN Clustering and 10 cross validation are used. Accuracy of 0.79, sensitivity of 0.82, and specificity = 0.78 is obtained. It is a Single- center retrospective study and only used non-contrast enhanced CT images. Šutić et al.[27] discussed prominent diagnostic and prognostic biomarkers for lung cancer. New biomarkers were identified for Non-Small Cell Lung Cancer. Terada et al.[28] built a deep learning model to predict the presence of ALKr. AUC of 0.73 is obtained. Further study should be addressed to improve accuracy of ALKr prediction. Tripathi et al.[29] used neural network model for predicting three major gene mutations in non-small cell lung cancer. Those 3 gene mutations are epidermal growth factor receptor (EGFR), Kirsten rat sarcoma virus (KRAS), and Anaplastic lymphoma kinase (ALK). Efficient Nets performs well for the prediction of gene mutation in lung cancer. Tsou and Wu[30] used deep learning techniques on histopathology images from TCGA to categorize PTCs (papillary thyroid carcinoma) into BRAF or RAS mutations. Area under the curve of 0.878–0.951 is obtained. Accuracy of prediction for BRAF group is 90.9% and 100% for RAS group. BRAF PTC was not considered as a homogenous group in clinical studies. The work did not consider the whole-slide images and expertise is required for selecting the region of interest in the slides. Prediction of EGFR mutation and PD-L1 expression without any invasive method was explored in the work[31]. 3D convolutional neural network and 5 cross validation is used. AUCs of 0.96, 0.76, and 0.76 in the training, validation, and test cohorts are obtained. This model has an accuracy of 0.90, a sensitivity of 0.74 and a specificity of 0.93 in the training set for the overall four- way classification. Single-center retrospective study and only limited to EFGR and PD-L1. Wang et al.[32] explained the significance of breath omics testing which is used for the detection and screening for lung cancer. Breath biomarkers are identified using HPPITOFMS. Yang et al.[33] prediction of recurrence and survivability of different types of lung cancer was explored with genomic, clinical, and demographic data along with copy number variation (CNV) and mutation information of 15 selected genes. Three common ML methods (decision trees, neural networks and support vector machines) were built, and their performance was compared. The best predictive results were obtained using CART Tree model. Yu et al.[34] explored CT image features and ML models for prediction of pathologic stage in NSCLC. Xie et al.[35] identified diagnostic biomarkers for lung cancer by combining metabolomics and machine learning methods. The AUC is 0.989, the sensitivity is 98.1%, and the specificity is 100%. Metabolic biomarkers are very useful for early detection of lung tumors.

Few studies are limited to classification of one or two biomarkers only. There is misclassification of biomarkers in a few cases. Very few studies deal with histopathology images as they are very challenging. Histopathological images are noisy with features of different unrelated cancers or normal tissues.

3. Materials and methods

The histopathology images are obtained from The Cancer Genome Atlas (TCGA). Histopathology images are examination of small piece of tissues by microscope. They are downloaded from the TCGAportal[36].

Genetic Mutation data contains 40,544 identifiers X 513 samples. Gene mutations data is given as input to the model in label file. Gene mutation data is downloaded from GDC website[37]. **Table 1** gives an overview of the gene mutation data.

Table 1. Description of mutation data.

Number of records	Number of features	Number of classes	Class description
40544	513	Class 1	Biomarker is Present
		Class 0	Biomarker is not Present

The proposed system is diagrammatically depicted in **Figure 1**. Tiles and genetic biomarkers (labels) are used to train the model which predicts the genetic mutations in a tile. Label file denotes the biomarkers present in a tilefor training the deep learning model. This helps the model to identify mutations, present in a tile.

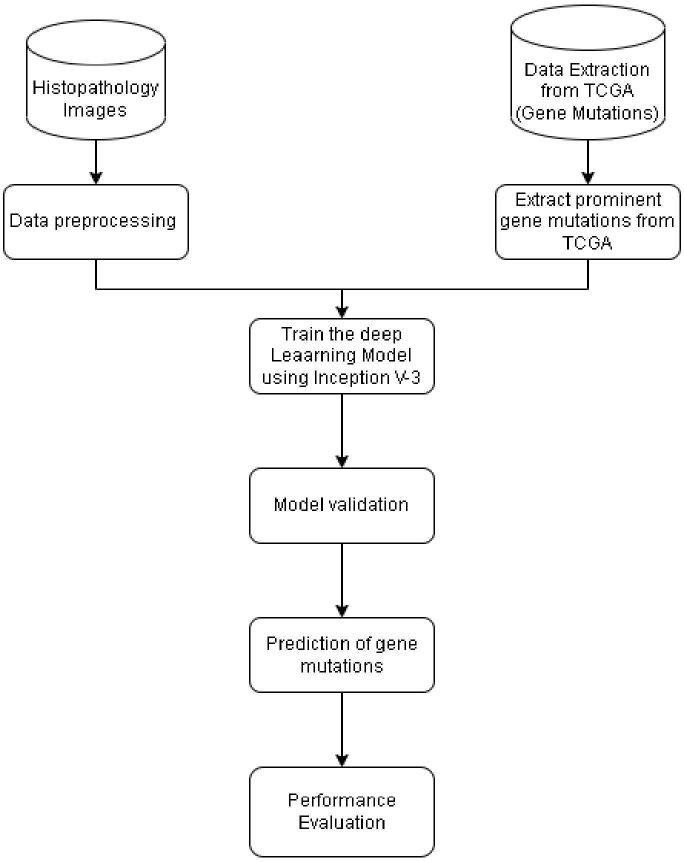

Histopathology Images
Data Extraction from TCGA (Gene Mutations)
Data preprocessing
Extract prominent gene mutations from TCGA
Train the deep Leaarning Model using Inception V-3
Model validation
Prediction of gene mutations
Performance Evaluation

Figure 1. Model for predicting probability of gene mutation.

3.1. Data extraction

Lung Adenocarcinoma (non-small cell lung cancer) is the most commonly occurring lung cancer type. It falls under the category of Non-Small Cell Lung Cancer (NSCLC). Lung cancer arises from the mucosal glands and represents 40% of lungs. Except stage 1, all other stages have worse prognosis.

As depicted in **Figure 2**, image dataset is obtained from TCGA and separate the dataset into training set, validation set and testing set.

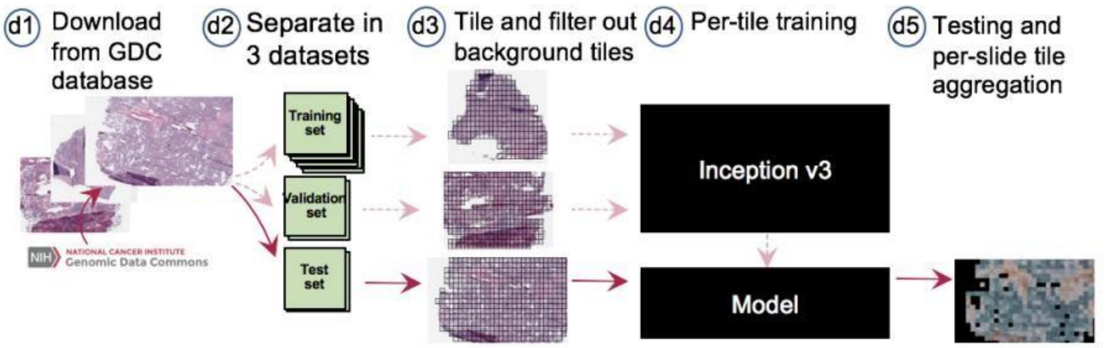

d1 Download from GDC database
d2 Separate in 3 datasets
d3 Tile and filter out background tiles
d4 Per-tile training
d5 Testing and per-slide tile aggregation
Training set
Validation set
Test set
Inception v3
Model
NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

Figure 2. Data extraction.

3.2. Data pre-processing

The process starts with downloading data from TCGA. SVS images are first tiled in non-overlapping 299 × 299 pixels window using Open slide library. The slides with low amount of information are removed. Tiles which occupied with more than 50% background were removed. Tiles are sorted into 3 categories training set, testing set and validation set. Jpg images are converted into Tensor Flow (TF) Records. The data pre-processing steps are depicted in **Figure 3**.

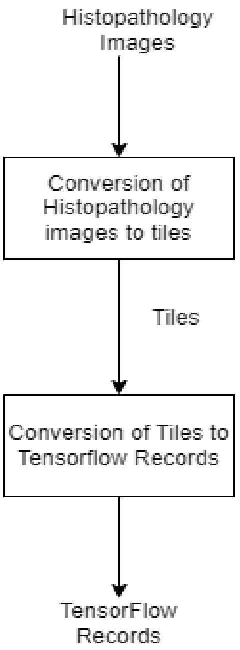

Histopathology Images
Conversion of Histopathology images to tiles
Tiles
Conversion of Tiles to Tensorflow Records
TensorFlow Records

Figure 3. Flow of data pre-processing.

Input: Histopathology Images Output: TensorFlow Records (TF Records).

3.3. Deep learning model

3.3.1. Convolutional Neural Network (CNN)

CNN architecture consists of a collection of convolutional filters (kernels). In CNN architecture maps input to output in an non-linear manner which adds complex learning nature to CNN. InceptionV3 model is used due to its best classification accuracy.

3.3.2. InceptionV3

- InceptionV3 architecture is going to be used to train the model. InceptionV3 is a 48 layer deep Convolutional neural network. The network has an image input size of 299 × 299. It is used for image recognition and image classification.
- The various layers in the InveptionV3 model are shown in the **Figure 4**. Batch normalization is used

extensively throughout the model and applied to activation inputs. Loss is computed using SoftMax.

- InceptionV3 is trained for training steps of 500.Separate validation and test datasets are used to reduce the risk of overfitting. **Figure 4** depicts the structure of InceptionV3 model.

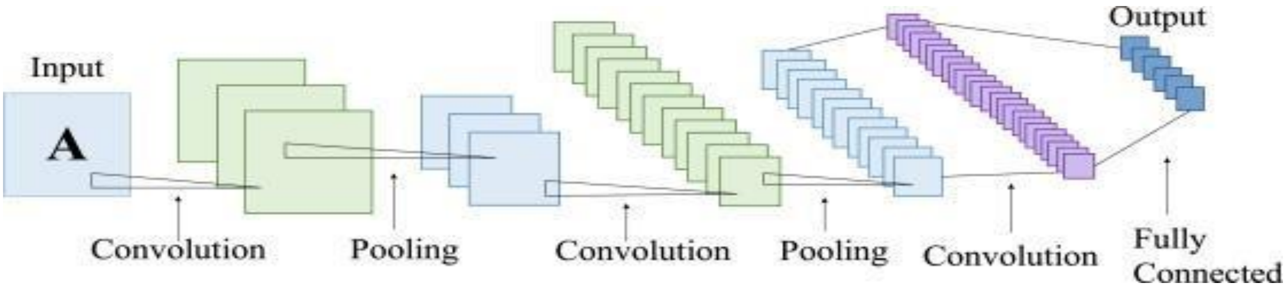


Figure 4. Inception model.

Figure 5 shows the overall flow of events in predicting the probability of gene mutations. The Tensor Flow records and corresponding genetic biomarkers are combined and used to train the InceptionV3 model which categorizes the prognostic and predictive biomarker and the probability of each gene mutated was obtained. The model was evaluated using performance metrics such as accuracy and cross entropy.

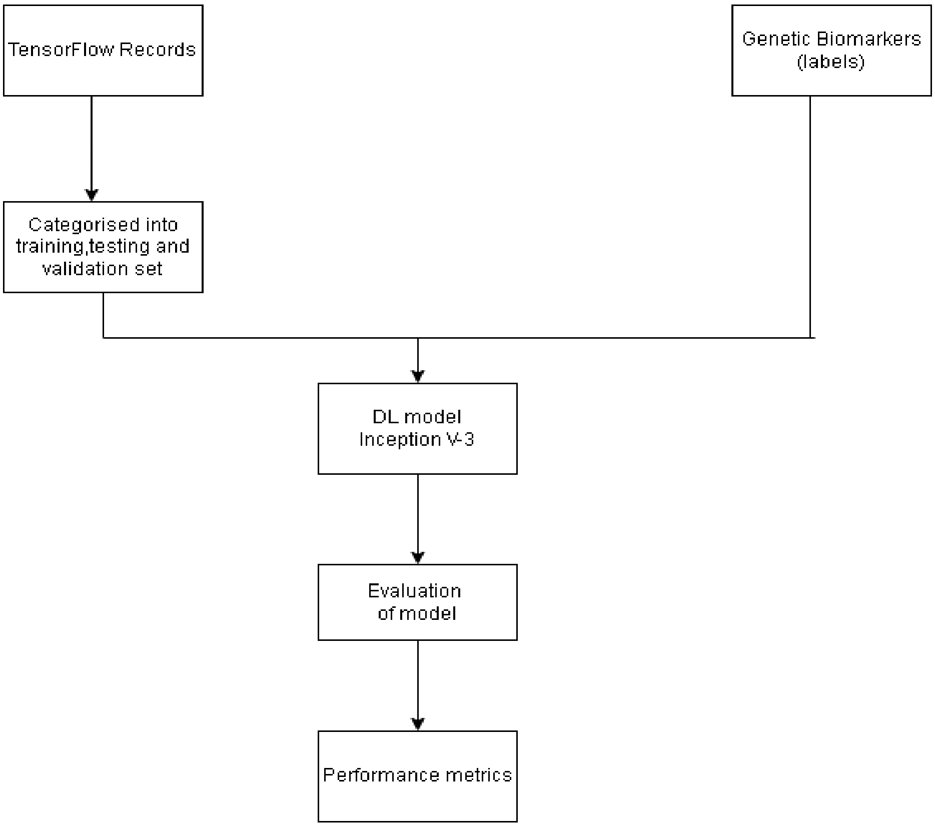

TensorFlow Records
Genetic Biomarkers (labels)
Categorised into training,testing and validation set
DL model Inception V-3
Evaluation of model
Performance metrics

Figure 5. Deep learning model.

Input: TensorFlow Records and Gene Mutations.

Output: Percentage of probability of occurrence of gene mutations.

3.4. Model evaluation

The following metrics are used to evaluate the performance of the models.

3.4.1. Accuracy

Accuracy is one metric for evaluating classification models. Accuracy is calculated using the following formula for multi-label classification. **Figure 6** shows the confusion matrix for computing accuracy.

$$\text{Accuracy} = TP + TN/(TP + TN + FP + FN)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

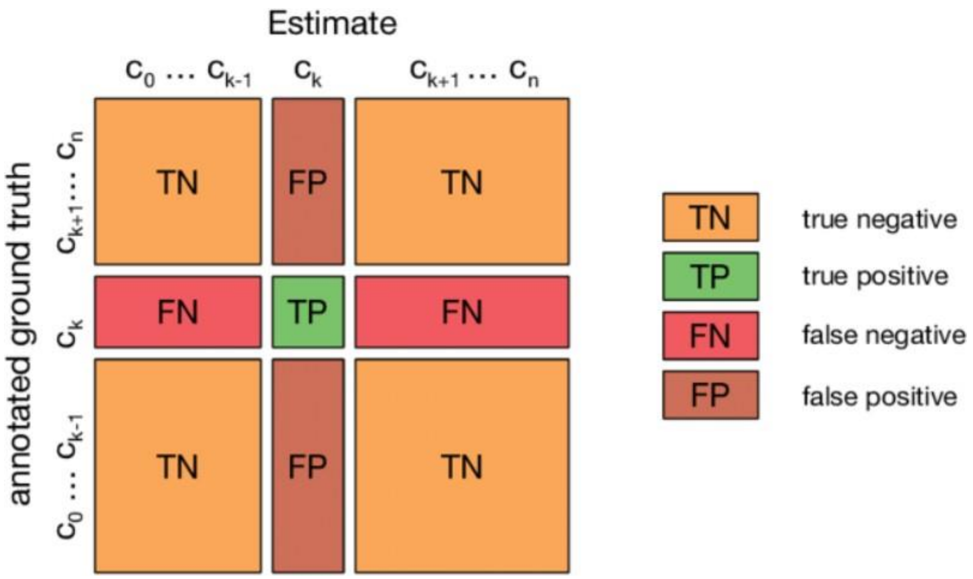


Figure 6. Confusion matrix for multi label classification.

3.4.2. Cross entropy

Cross-entropy between two probability distributions and over the same underlying set of events which measures the average number of bits needed to identify an event. It also represents the deviation of actual to the expected outcome. The lower the deviation, the better the performance of the model.

4. Results

InceptionV3 based model predicted gene mutations from histopathology images with an accuracy of 82.36% and cross entropy of 37.62%.

In **Figure 7**, accuracy of InceptionV3 for lung cancer can be observed. After 500 training steps, the accuracy of the model is 82.82% for training set and 74.95% is obtained for validation set.

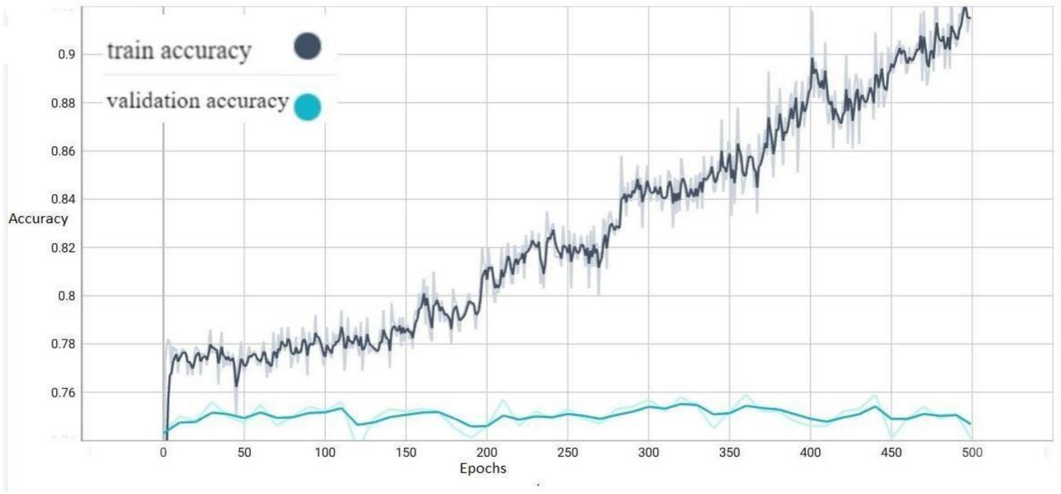


Figure 7. Accuracy of InceptionV3 for lung cancer.

In **Figure 8**, cross entropy of InceptionV3 for lung cancer can be observed. After 500 training steps, the cross entropy of the model is 36.9% for training set and 59.07% is obtained for validation set.

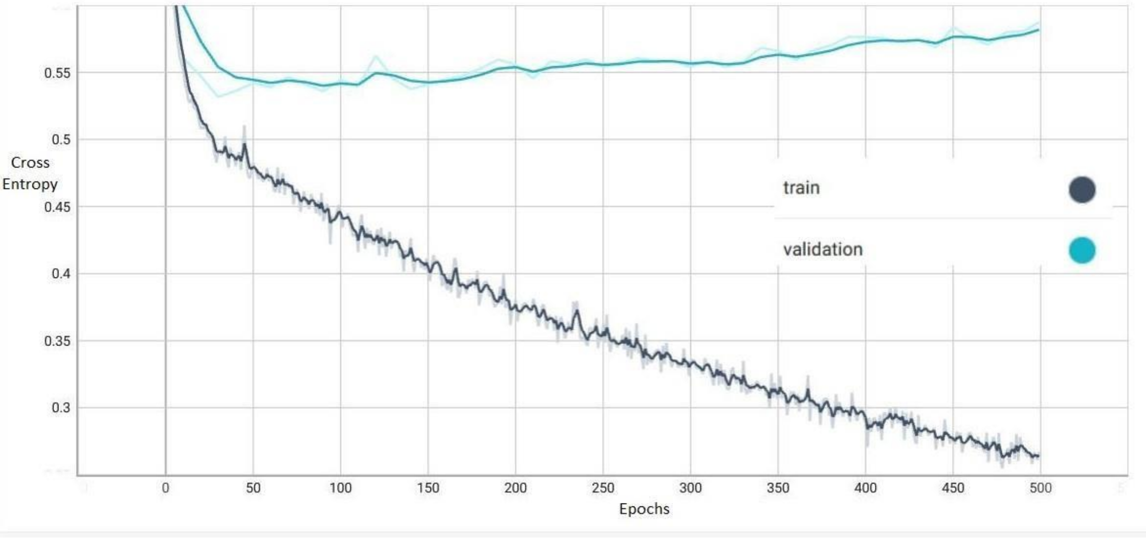
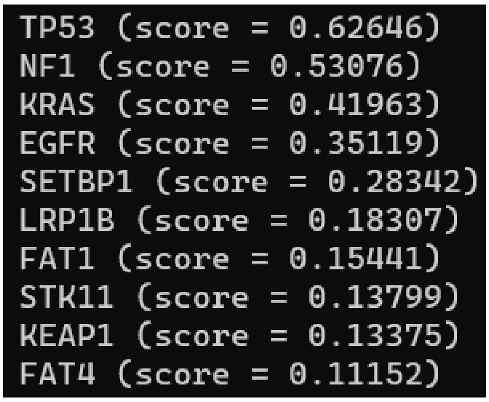


Figure 8. Cross entropy of InceptionV3 for lung cancer.

Figure 9 displays the probability of gene mutation obtained in a tile. Gene mutation along with score is displayed. In **Table 2** probability of occurrence of gene mutation along with category of biomarkeris tabulated.



```
TP53 (score = 0.62646)
NF1 (score = 0.53076)
KRAS (score = 0.41963)
EGFR (score = 0.35119)
SETBP1 (score = 0.28342)
LRP1B (score = 0.18307)
FAT1 (score = 0.15441)
STK11 (score = 0.13799)
KEAP1 (score = 0.13375)
FAT4 (score = 0.11152)
```

Figure 9. Probability of gene mutation in a tile.

Table 2. Probability of gene mutation along with category of biomarker.

Gene name	Probability of occurrence of gene mutation (in %)	Category of Biomarker
EGFR	33.94	Predictive Biomarker
KRAS	41.55	Predictive Biomarker
SETBP1	28.3	Prognostic Biomarker
NF1	54.3	Prognostic Biomarker
LRP1B	19.63	Predictive Biomarker
FAT1	15.09	Prognostic Biomarker
FAT4	11.18	Prognostic Biomarker
STK11	14.37	Prognostic Biomarker
KEAP1	13.37	Prognostic Biomarker
TP53	62.71	Predictive Biomarker

5. Discussion

To accurately predict genetic mutations from histopathological images, multiple deep learning techniques were explored. Post the literature survey, a few problems were identified. Firstly, the issue with large size of histopathology images was identified. Due to enormous file size, histopathology images are converted into tiles for applying deep learning techniques. **Figure 7** shows a sample tile. The model was trained using the tiles and label files. However, there are a few limitations. A small dataset has been used and a very few performance metrics has been applied to thetrained model. Some statistical tests like Wilcoxon and ANOVA are not directly applicable to evaluating the performance of deep learning models like InceptionV3. There are alternative approaches that are used to ensure the quality of the proposed method. It involves splitting the available data into training and validation sets and repeatedly training and evaluating the model on different subsets. Performance metrics like accuracy and cross entropy are applied on the proposed model to test its performance. **Figure 10** shows the image of a tile.

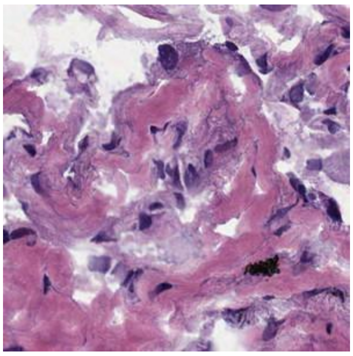


Figure 10. A sample image of Tile.

6. Conclusion

Gene mutation prediction from histopathology images using deep learning was attempted. Histopathology pathology images are converted into tiles using data preprocessing techniques. Deep Learning model InceptionV3 is trained with tiles and their respective label files where a label file denotes mutations in a particular image. Deep Learning Model was trained for training steps of 500. InceptionV3 predicted gene mutations from histopathology images with an accuracy of 82.36% and cross entropy of 37.62%. Different deep learning models like AlexNet and ResNet can also be used formutation prediction which can be explored further.

Author contributions

Conceptualization, LYV and GVNAV; methodology, GVNAV; software, GVNAV; validation, LYV, DVVP, GVNAV; formal analysis, LYV, DVVP; investigation, LYV, DVVP; resources, GVNAV and CV; data curation, GVNAV and CV; writing—original draft preparation, GVNAV and CV; writing—review and editing, LYV and GVNAV; visualization, GVNAV; supervision, LYV and DVVP; project administration, LYV and DVVP. All authors have read and agreed to the published version of the manuscript.

Declaration

The manuscript consists of sub sections such as Abstract, Introduction, Materials & Methods, Results, Discussion and Conclusion. The manuscript complies with the format as mentioned in the instructions to the authors. The compliance of Ethical Standards was mentioned below.

Ethical responsibilities of authors

The manuscript is not submitted to more than one journal for simultaneous consideration.

Ethical approval

This article does not contain any studies of human participants or animals performed by any of the authors.

Informed consent

Informed consent is not necessary as this article does not involve human or animal participants.

Conflict of interest

The authors declare no conflict of interest.

References

1. Uludag U, Pankanti S, Prabhakar S, Jain AK. Biometric cryptosystems: Issues and challenges. *Proceedings of the IEEE* 2004; 92(6): 948–960. doi: 10.1109/jproc.2004.827372
2. Thukral R, Aggarwal AK, Arora AS, et al. Artificial intelligence-based prediction of oral mucositis in patients with head-and-neck cancer: A prospective observational study utilizing a thermographic approach. *Cancer Research, Statistics, and Treatment* 2023; 6(2): 181–190. doi: 10.4103/crst.crst_332_22
3. Kumar A. Light propagation through biological tissue: Comparison between Monte Carlo simulation and deterministic models. *International Journal of Biomedical Engineering and Technology* 2009; 2(4): 344. doi: 10.1504/ijbet.2009.027798
4. Aggarwal AK. Rehabilitation of the blind using audio to visual conversion tool. *British Journal of Healthcare and Medical Research* 2014; 1(4): 24–31. doi: 10.14738/jbemi.14.395
5. Kaur A, Chauhan APS, Aggarwal AK. Machine learning based comparative analysis of methods for enhancer prediction in genomic data. In: Proceedings of the 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT); 28–29 September 2019; Jaipur, India. pp. 142–145.
6. Maini S, Aggarwal AK. Camera position estimation using 2D image dataset. *International Journal of Innovations in Engineering and Technology (IJIET)* 2018; 10(2): 199–203. doi: 10.21172/ijict.102.29
7. Thukral R, Kumar A, Arora AS. Effects of different radiations of electromagnetic spectrum on human health. In: Proceedings of the 2020 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS); 22–23 February 2020; Bhopal, India. pp. 1–6.
8. Amini M, Hajianfar G, Avval AH, et al. Overall survival prognostic modelling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonised radiomics features: The quest for the optimal machine learning algorithm. *Clinical Oncology* 2022; 34(2): 114–127. doi: 10.1016/j.clon.2021.11.014
9. Bhardwaj P, Raipuria G, Bhatt N, Singhal N, Joshi UY, Kondragunta C. A deep learning method for tumour region identification and tumour proportion score estimation of PD-L1 expression in non-small cell lung carcinoma. *Journal of Pathology Informatics* 2022; 13: 100041. doi: 10.1016/j.jpi.2022.100041.
10. Chang C, Sun X, Wang G, et al. A machine learning model based on PET/CT radiomics and clinical characteristics predicts ALK rearrangement status in lung adenocarcinoma. *Frontiers in Oncology* 2021; 11: 603882. doi: 10.3389/fonc.2021.603882
11. Chen M, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precision Oncology* 2020; 4(1): 14. doi: 10.1038/s41698-020-0120-3.
12. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* 2020; 1(8): 800–810. doi: 10.1038/s43018-020-0085-8
13. Gao Y, Zhou R, Lyu Q. Multiomics and machine learning in lung cancer prognosis. *Journal of Thoracic Disease* 2020; 12(8): 4531–4535. doi: 10.21037/jtd-2019-itm-013
14. Korpanty GJ, Graham DM, Vincent MD, Leighl NB. Biomarkers that currently affect clinical practice in lung cancer: EGFR, ALK, MET, ROS-1, and KRAS. *Frontiers in Oncology* 2014; 4: 204. doi: 10.3389/fonc.2014.00204
15. Mayer C, Ofek E, Fridrich DE, et al. Direct identification of ALK and ROS1 fusions in non-small cell lung cancer from hematoxylin and eosin-stained slides using deep learning algorithms. *Modern Pathology* 2022; 35(12): 1882–1887. doi: 10.1038/s41379-022-01141-4
16. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 2015; 13: 8–17. doi: 10.1016/j.csbj.2014.11.005
17. Morgado J, Pereira T, Silva F, et al. Machine learning and feature selection methods for EGFR mutation status prediction in lung cancer. *Applied Sciences* 2021; 11(7): 3273. doi: 10.3390/app11073273
18. Murchan P, Ó’Brien C, O’Connell S, et al. Deep learning of histopathological features for the prediction of tumour molecular genetics. *Diagnostics* 2021; 11(8): 1406. doi: 10.3390/diagnostics11081406
19. Ni M, Liu X, Wu J, et al. Identification of candidate biomarkers correlated with the pathogenesis and prognosis of non-small cell lung cancer via integrated bioinformatics analysis. *Frontiers in Genetics* 2018; 9: 469. doi: 10.3389/fgene.2018.00469
20. Patil SS, Shetty D, Pawar VS. Novel machine learning algorithm for prevalent gene biomarkers for effective cancer treatment by detecting its PH. *Computer Science & Information Technology* 2021; 11(7): 155–170. doi: 10.5121/csit.2021.110713

21. Roman-Canal B, Moiola CP, Gatius S, et al. EV-associated miRNAs from pleural lavage as potential diagnostic biomarkers in lung cancer. *Scientific Reports* 2019; 9(1): 1–9. doi: 10.1038/s41598-019-51578-y
22. Santarpia M, Liguori A, D'Aveni A, et al. Liquid biopsy for lung cancer early detection. *Journal of Thoracic Disease* 2018; 10(S7): S882–S897. doi: 10.21037/jtd.2018.03.81
23. Seijo LM, Peled N, Ajona D, et al. Biomarkers in lung cancer screening: Achievements, promises, and challenges. *Journal of Thoracic Oncology* 2019; 14(3): 343–357. doi: 10.1016/j.jtho.2018.11.023
24. Shiri I, Amini M, Nazari M, et al. Impact of feature harmonization on radiogenomics analysis: Prediction of EGFR and KRAS mutations from non-small cell lung cancer PET/CT images. *Computers in Biology and Medicine* 2022; 142: 105230. doi: 10.1016/j.compbiomed.2022.105230
25. Song L, Zhu Z, Mao L, et al. Clinical, conventional CT and radiomic feature-based machine learning models for predicting ALK rearrangement status in lung adenocarcinoma patients. *Frontiers in Oncology* 2020; 10: 369. doi: 10.3389/fonc.2020.00369
26. Silva F, Pereira T, Morgado J, et al. EGFR assessment in lung cancer CT images: Analysis of local and holistic regions of interest using deep unsupervised transfer learning. *IEEE Access* 2021; 9: 58667–58676. doi: 10.1109/access.2021.3070701
27. Šutić M, Vukić A, Baranašić J, et al. Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (NSCLC) management. *Journal of Personalized Medicine* 2021; 11(11): 1102. doi: 10.3390/jpm11111102
28. Terada Y, Takahashi T, Hayakawa T, et al. Artificial intelligence—Powered prediction of alk gene rearrangement in patients with non-small-cell lung cancer. *JCO Clinical Cancer Informatics* 2022; 6: e2200070. doi: 10.1200/cci.22.00070
29. Tripathi S, Moyer EJ, Augustin AI, et al. RadGenNets: Deep learning-based radiogenomics model for gene mutation prediction in lung cancer. *Informatics in Medicine Unlocked* 2022; 33: 101062. doi: 10.1016/j.imu.2022.101062
30. Tsou P, Wu CJ. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: Applying a deep convolutional neural network. *Journal of Clinical Medicine* 2019; 8(10): 1675. doi: 10.3390/jcm8101675
31. Wang C, Xu X, Shao J, et al. Deep learning to predict EGFR mutation and PD-L1 expression status in non-small-cell lung cancer on computed tomography images. *Journal of Oncology* 2021; 2021: 5499385. doi: 10.1155/2021/5499385
32. Wang P, Huang Q, Meng S, et al. Identification of lung cancer breath biomarkers based on perioperative breathomics testing: A prospective observational study. *EClinicalMedicine* 2022; 47: 101384. doi: 10.1016/j.eclinm.2022.101384
33. Yang Y, Xu L, Sun L, et al. Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal* 2022; 20: 1811–1820. doi: 10.1016/j.csbj.2022.03.035
34. Yu L, Tao G, Zhu L, et al. Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. *BMC Cancer* 2019; 19(1): 1–12. doi: 10.1186/s12885-019-5646-9
35. Xie Y, Meng WY, Li RZ, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology* 2021; 14(1): 100907. doi: 10.1016/j.tranon.2020.100907
36. Histopathology images from TCGA. Available online: https://portal.gdc.cancer.gov/projects/TCGA-LUAD (accessed on 24 March 2023).
37. Gene mutation data from GDC. Available online: https://xenabrowser.net/datapages/ (accessed on 25 March 2023).