How to Cite:
Anita, C. S., Vasukidevi, G., Rajalakshmi, D., Selvi, K., & Ramesh, T. (2022). Lung cancer prediction model using machine learning techniques. *International Journal of Health Sciences*, *6*(S2), 12533–12539. https://doi.org/10.53730/ijhs.v6nS2.8306

Lung cancer prediction model using machine learning techniques

C. S. Anita
Professor, Department of AIML, R.M.D. Engineering College

Vasukidevi.G
Assistant Professor, Department of Science & Humanities, R.M.K.College of Engineering and Technology

D. Rajalakshmi
Associate Professor, Department of CSE, R.M.D. Engineering College

K. Selvi
Professor, Department of CSE, R.M.K. Engineering College

Ramesh. T.
Associate Professor, Department of CSE, R.M.K. Engineering College

***Abstract*---**Lung cancer is cancer that forms in tissues of the lung, usually in the cells that line the air passages. It is the leading cause of cancer death in both men and women. Some of the Symptoms are Chest pain or discomfort, Trouble breathing, Wheezing, Blood in sputum (mucus coughed up from the lungs),Hoarseness, Loss of appetite, etc. Sometimes lung cancer does not cause any signs or symptoms. It may be found during a chest x-ray done for another condition. So early prediction of disease is very important to avoid death. So many machine learning algorithms are used to predict the lung cancer early but lack of accuracy. To overcome disease prediction accuracy issues, Gaussian Naive Bayes machine learning algorithm is used. The performance of the proposed GNB algorithm is evaluated using UCI Machine Learning Repository. The performance analysis shows GNB prediction model achieves 97.5%.

***Keywords*---**lung cancer, GNB, UCI dataset prediction model, accuracy.

Introduction

lung cancer, also known as lung carcinoma,[8] since about 98–99% of all lung cancers are carcinomas, is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung.[9] Lung carcinomas derive from transformed, malignant cells that originate as epithelial cells, or from tissues composed of epithelial cells. Other lung cancers, such as the rare sarcomas of the lung, are generated by the malignant transformation of connective tissues (i.e. nerve, fat, muscle, bone), which arise from mesenchymal cells. Lymphomas and melanomas (from lymphoid and melanocyte cell lineages) can also rarely result in lung cancer. In time, this uncontrolled growth can spread beyond the lung – either by direct extension, by entering the lymphatic circulation, or via the hematogenous, bloodborne spread – the process called metastasis – into nearby tissue or other, more distant parts of the body.[10] Most cancers that start in the lung, known as primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, and chest pains.

The vast majority (85%) of cases of lung cancer are due to long-term tobacco smoking. About 10–15% of cases occur in people who have never smoked.[11] These cases are often caused by a combination of genetic factors and exposure to radon gas, asbestos, second-hand smoke, or other forms of air pollution.[12][13] Lung cancer may be seen on chest radiographs and computed tomography (CT) scans.[14] The diagnosis is confirmed by biopsy, which is usually performed by bronchoscopy or CT-guidance.[15].

Figure 1: Lung Cancer Statistics(2021)

Related works

In 2019, Moradi et al. [16] compared different techniques to differentiate lung cancer nodules from nonnodules. To reduce/eliminate the false positive predictions they have come up with 3D Convolutional Neural Network Technique. Nodules exist in different sizes and using just one CNN can result in false detections. So they divided the nodules into four groups according to their size. And they have used four different sizes of 3D CNN. They combined all those 4 classifiers to get better results. Each CNN consists of a nubmber of 3D CNN which are all varying sizes. All 4 classifiers were combined in order to produce results which were better.

In 2019, Ruchita Tekade et al. [17]. proposed a method using 2 architectures, one for the segmentation of nodules and the second one to determine the malignancy level. For determining the malignancy level CNN is used for classification as well as for the feature extraction, max pooling is used for sub pooling, ReLU as the activation function, and softmax is the classifier used to perform the classification and assign malignancy level. Adam classifier is used to optimize weight selection in convolutional kernels. For the segmentation of CT scanned images, pre-processing is done using simple thresholding, clear border, morphology erosion, morphology closing, and morphology opening respectively. Using U-Net segmentation masses are generated for lung CT scan images and lung nodules are segmented. This experiment was conducted on LIDC-IDRI, LUNA16, and Data Science Bowl2017 datasets. This approach gives an accuracy of 95.66% and loss 0.09 and dice coefficient of 90%and for predicting log loss 38% using U-Net to segment and further predict malignancy levels

In 2018, Margarita Kirienko et al. [18] suggested a CNN-based approach with 69%, 69%, and 87% accuracy invalidation, test, and training sets respectively. Tumour, Node, Metastasis (TNM) staging was used to stage lung cancer from 1 to 4. Fluorode oxyglucose positron emission tomography (FDG-PET)/ Computed Tomography (CT) images were used as input. These images were classified into either T1- T2 or T3-T4 using CNN. The system was developed using two networks - a classifier and a feature extractor. The feature extractor was used for relevant features that are to be extracted and a classifier was used to classify the patch. The experiment was performed on 472 patients (T1-T2 = 353 and T3-T4 = 119). In 2020, QINGHAI ZHANG et al. [19] proposed a method for designing of Lung nodule detection system which is automatic. The dataset used for the proposed method is LIDC-IRDI public dataset. The proposed method used for this study is Multi-Scene Deep Learning Framework which contains several steps. CT images are given as input and the probability distribution of distinct gray levels is obtained by threshold segmentation that is Histogram. Correcting the smooth lung outlines is the main aim for the lung parenchyma segmentation process. The replacement of the vein system in the lung helps to identify the nodule structure. Vessel filters are used for removing the vessels which reduce the number of false positive. The design of CNN contains a pooling layer, a convolutional layer, and a fully integrated layer. Segmentation and classification identify Class 1 and Class2 that are two class of image data and discrete images which are separated from the lung images respectively.

Proposed System

The proposed Lung cancer prediction model consists of four different steps such as image pre-processing, image segmentation, feature extraction and image classification. They need to be well pre-processed before the actual use. Various Image pre-processing techniques are used to discard noise and to make images suitable for use. This helps in the betterment of the performance of the whole system and hence the accuracy. The method of partitioning an image into several segments is known as image segmentation. Segmentation of image is done majorly to find boundaries in the given image. The process of analyzing the image becomes easier as segmentation reduces the image complexity. Feature Extraction is a method by which we aim at reducing the number of dimensions

that our raw data contains so that it is easier to process and is in a form of manageable classes. Variables in a huge number requiring computational resources in order to process and produce results is characteristic for the massive amounts of data. Classification of images is a basic task that seeks to interpret a picture as a whole. By assigning it to a particular label, the purpose is to identify the image. Image Classification usually refers to images where only one object appears and is examined.

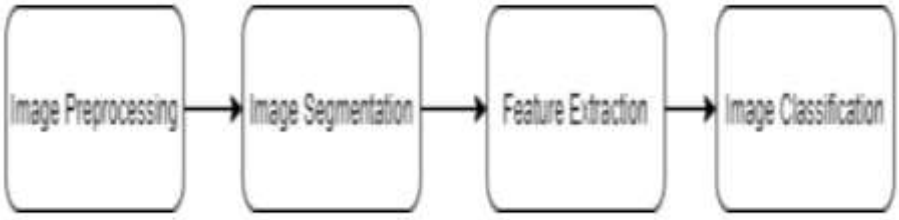

Figure 2: Proposed Prediction System Model

Dataset Description

The dataset description are shown in table

Feature name	Values	Remarks
Gender	Male (M),Female(F)	
Age	Range: 25-45 years=1 and Range 46-65 years =0	Age in years
Smoker	Yes=1 or No=2	
tumor location	Central Location =1,Peripheral Location=0	
t-stage,	Yes=1, no=0	t represents the main tumor
n-stage	Yes=1, no=0	spread to neighboring lymph nodes
Stage	2,3	2nd stage,3rd stage
Timing	More than a year=1 or Less than a year=0	Survival time in years (count in days)
Diabetes	Yes=1, no=0	
Status	Censored=1, dead=0	censoring status
meal.cal	More=1,Less=0	Calories consumed at meals
wt.loss	Decreased=1 , Increased=0	Weight loss in last six months
ph.ecog	5=dead , 0=good	ECOG performance score
ph.karno	Good=80-100: Able to carry on normal activity and to work rated by physician. Average=50-70: Unable to work Bad=0-60: Unable to care for self& disease may be progressing rapidly.	Karnofsky performance score index is an assessment tool for functional therapies and to assess the prognosis in individual patients.

Table 1: Dataset Description

Experimental Analysis

The proposed lung cancer prediction model is implemented with following experimental environment.

Table 2 : Implementation and experimental environment

Item	Specification(s)
Programming language	Python Keras with TensowrFlow 2
Operating System	Windows 10
CPU	Intel® CORE i7-2.8GZ
Memory	16 GB
Graphics Processor	Intel® UHD Graphics 630

Evaluation Metrics

The effectiveness of the proposed technique is demonstrated using different evaluation metrics, by measuring the true and/or misclassification of lung cancer positive/negative cases in the X-ray images (i.e., testing dataset). These metrics have been directly driven from the confusion matrix illustrated below. The performance of Lung Cancer prediction model can be evaluated using confusion matrix. A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

Table 3: Prediction Class

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Table 4: Lung Cancer prediction

ML Techniques	Accuracy	Precision	Sensitivity	Specificity
SVM	78	87	78	87
RF	87	88	89	88
NB	88	89	90	89
ANN	89	90	92	90
GNB	98	92	97	98

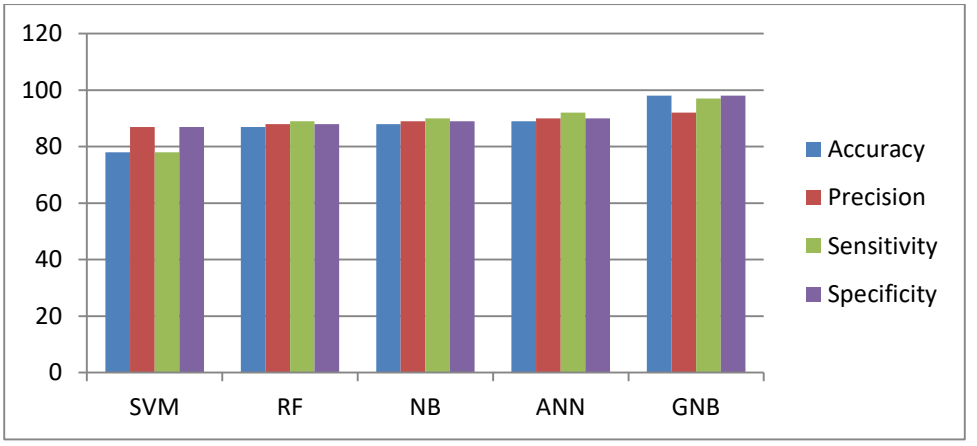

Figure 3: Lung Cancer Prediction

Conclusion

Cancer is a disease in which cells in the body grow out of control. When cancer starts in the lungs, it is called *lung cancer.* Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body, such as the brain. Cancer from other organs also may spread to the lungs. When cancer cells spread from one organ to another, they are called *metastases .In this work, GNB machine leaning technique are used to predict the lung cancer.* The performance of the proposed GNB algorithm is evaluated using UCI Machine Learning Repository. The performance analysis shows GNB prediction model achieves 98% accuracy compare to other machine learning techniques.

References

1. Fan, Jianqing; Han, Fang; Liu, Han (2014-06-01):"Challenges of Big Data analysis: National Science Review". 1 (2): 293–314. ISSN 2095-5138. PMC 4236847. PMID 25419469. doi:10.1093/nsr/nwt032.
2. Tina M. St. John M.D. (2005).:"With Every Breath: A Lung Cancer Guidebook" .1(1):75-82. ISBN 0-9760450-2-8,www.lungcancerguidebook.org.
3. B. Sobolev, A. Levy, and S. Goring, Eds (2016).: "Health Services Data: Big Data Analytics for Deriving Predictive Healthcare Insights, inData and Measures in Health Services Research". SpringerUS, 2016, 1(1)1–17,DOI: 10.1007/978-1-4899-7673-4_2-1,ISBN 978-1-4899-7673-4.

4. http://www.news-medical.net/news/20120530/Insulinuse-linked-to-lung-cancer-risk-in-diabetes.aspx (2012)
5. Dutkowska, Adam Antczak. (2016).: Comorbidities in lung cancer , AgataEwa,Pneumonologia i AlergologiaPolska 84 (3): 186–192.
6. Xie X, Liu Q, Wu J, Wakui M.(2009),: “Impact of cigarette smoking in type 2 diabetes development”, ActaPharmacologicaSinica.30(6):784-787. doi:10.1038/aps.2009.49.
7. Chang SA.(2012):Smoking and Type 2 Diabetes Mellitus”. Diabetes & Metabolism,Journal;36(6):399-403 doi:10.4093/dmj.2012.36.6.399.
8. Maddatu, Judith et al.(2017).:”Smokingand the risk of type 2 diabetes”. Translational Research:184(1),101-107.http://dx.doi.org/10.1016/j.trsl.2017.02.004.
9. Appari A, Eric Johnson M, Anthony DL.(2013):”Meaningful use of electronic health record systems and process quality of care: evidencefrom a panel data analysis of U.S. acute-care hospitals”. Health Serv Res.48(1):354–75.
10. Fitzhenry F, Murff HJ, Matheny ME, et al.(2013).:“Exploring the frontier of electronic health record surveillance: the case of postoperativecomplications”. Med Care51:509–16.
11. J.R. Quinlan.(1994).:“C4.5 programs for machine learning”. Morgan Kaufmann Publishers,(16):235-240.
12. Vapnik, V.(1995).:”Support-vector networks".Machine Learning. 20 (3): 273–297. doi:10.1007/BF00994018.
13. G. Dimitoglou, J. A. Adams, and C. M. Jim.(2012).: “Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of LungCancer Survivability”, CoRR, 4(8):1–9.
14. Hamid KarimKhani Z and et.al.(2015).:”A comparative survey on data mining techniques for breast cancer diagnosis and prediction-Survey”.Indian Journal of Fundamental and Applied Life Sciences.5 (S1): 4330-4339 ISSN: 2231–6345.
15. Olusayo D. Fenwa, Funmilola A. Ajala and AdebisiA.(2015).:”Adigun Classification of Cancer of The Lungs Using SVM andANN”.International Journal Of Computers &Technology.15 (1):2277-3061
16. Moradi P and Jamzad M 2019 Detecting Lung Cancer Lesions in CT Images using 3D Convolutional Neural Networks 4th Int. Conf. on Pattern Recognition and Image Analysis (IPRIA) pp. 114-118.
17. Tekade R and Rajeswari K 2018 Lung cancer detection and classification using deep learning. Fourth Int. Conf. on Computing Communication Control and Automation (ICCUBEA) pp. 1-5.
18. Vamsidhar Enireddy, R ,P Shobha Rani ,Anitha, Sugumari Vallinayagam, T Maridurai, T Sathish, E Balakrishnan "Prediction of human diseases using optimized clustering techniques" 2021Materials Today: ProceedingsVolume 46Pages 4258-4264 PublisherElsevier
19. Sasikala S, Bharathi M, Sowmiya BR 2018 Lung Cancer Detection and Classification Using Deep CNN International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2278-3075.
20. Liu Z, Yao C, Yu H and Wu T 2019 Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things Future Generation Computer Systems pp 1-9