IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIAGNOSE OF LUNG CANCER PREDICTION USING ENSEMBLE LEARNING

Dr. A. Gnana Soundari M.Tech., Ph.D , Christo Sujana.C

Associate Professor, Information Technology, Jeppiaar engineering college, Sholinganallur, Chennai, India

UG Student, Jeppiaar engineering college, Sholinganallur, Chennai, India

ABSTRACT-

The aim is to predict machine learning based techniques for lung cancer prediction. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will bedone on the entire given dataset. To propose a machine learning-based method toaccurately predict the lung cancer using supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface of lung cancer prediction by attributes.

KEYWORDS-

Dataset, Machine Learning Classification method, Python, Prediction of accuracy result.

I. INTRODUCTION

Lung cancer is due to uncontrollable growth of cells in the lungs. It causes a serious breathing problem in both inhale and exhale part of chest. Cigarette smoking and passive smoking are the principal contributor for the cause of lung cancer as per world health organization. The mortality rate due to lung cancer is increasing day by day in youths as well as in old persons as compared to other cancers.

This helps all others department to carried out other formalities. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code.

The following Involvement steps are,

- Define a problem
- Preparing data
- Evaluating algorithms
- Improving results
- Predicting results

The aim is to predict machine learning based techniques for lung cancer prediction. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variateanalysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will bedone on the entire given dataset.

II. EXISTING SYSTEM

Detecting lung nodules with low-dose computed tomography (CT) can predict the future risk suffering from lung cancers. There are a few studies on lung nodules with low-dose CT and detecting rate is very low at present. In order to accurately detect lung nodules with low-dose CT, this paper proposes a solution based on an integrated deep learning algorithm. The CT images are preprocessed via image clipping, normalization and segmentation, and the positive samples areexpanded to balance the number of positive and negative samples. The features of candidate lung nodule samples are learned by using convolutional neural network and residual network, and then import into long short-term memory network, respectively. We then fuse these features, continuously optimize the

network parameters during the training process, and finally obtain the model withan optimal performance. The experimental results prove that compared to other algorithms, all metrics in the proposed algorithm are improved. This model has an obvious anti-interference ability. It is stable and can identify lung nodules effectively, which is expected to provide auxiliary diagnostic for early screeningof lung cancers.

- They are using CT images and large amount of data required to classifyaccurately and it take a lot of time to train the model.
- Any modification made takes lots of effort to change the model.

III. PROPOSED SYSTEM

Machine learning supervised classification algorithms will be used to give datasetand extract patterns, which would help in predicting the likely patient affected or not, thereby helping them for making better decisions in the future.

Data collection

The data set collected for predicting the network attacks is split into Training setand Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using by ensemble learning model are applied on the Training set and based on the test result accuracy, Test set prediction is done.

- It improves accuracy score by comparing popular machine learning algorithms.
- These reports are to the investigation of applicability of machine learning techniques for detecting cancer in operational conditions by attribute prediction.

IV. REQUIREMENT SPECIFICATION

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

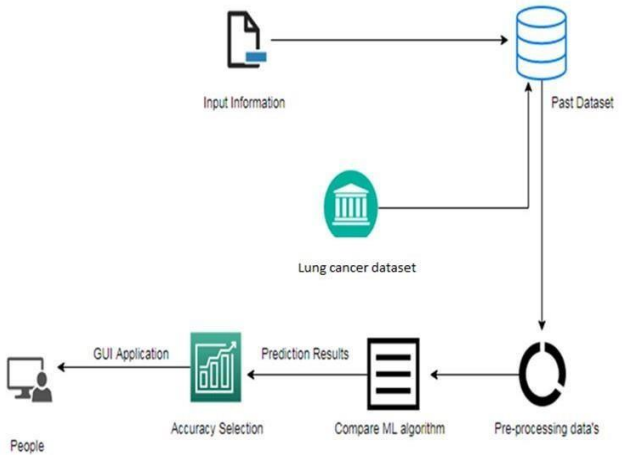
Functional requirements

Non-Functional requirements

Environment requirements

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirementsanalysis process. It lists requirements of a particular software system. The following details to follow the special libraries like NumPy, matplotlib and OpenCV.

V. SYSTEM ARCHITECHTURE



Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true errorrate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of dataused to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill onthe validation dataset is incorporated into the model configuration. The validationset is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list.

Data Validation/ Cleaning/Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values,duplicate values. A validation dataset is a sample of data held back from trainingyour model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the rawdata into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning methodof the data has to be in a proper manner. Some specified Machine Learning modelneeds information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set.

Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data.Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a datasetand can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization andexploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such aswith charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizingdata in Python and how to use them to better understand your own data.

Logistic Regression:

It is a statistical method for analysing a data set in which there are one or moreindependent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).
In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variableshould represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is,the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

Decision Tree:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

Support Vector Machines (SVM):

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a highpredictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to

method for a high-performing algorithm with little tuning.

- How to disentangle the many names used to refer to support vector machines.
- The representation used by SVM when the model is actually stored on disk.
- How a learned SVM model representation can be used to make predictions fornew data.
- How to learn an SVM model from training data.
- How to best prepare your data for the SVM algorithm.
- Where you might look to get more information on SVM.

RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learningalgorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting ina *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

K-Nearest Neighbor (KNN):

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance- weighted nearest neighbor algorithm, it weights the contribution of each of the kneighbors according to their distance using the following query giving greater weight to the closest neighbors. Usually, KNN is robust to noisy data since it is averaging the k-nearest neighbors.The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to labelother points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k "closest" instances. "Closeness" is determinedusing a proximity measurement (Euclidean) across all features.

Naive Bayes algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumptionbut results in a fast and effective method. The probability of a class value givena value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a predictionwe can calculate probabilities of the instance belonging to each class and select the class value with the highest probability Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to createwindows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants. There will be two kinds of pages like registration user purpose and login entry purpose of users.

Accuracy calculation:

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes.

E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no.

E.g. if actual class value indicates that this passenger survived and predicted classtells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

GUI

A graphical user interface (GUI) is a user interface built with graphical objects such as buttons, text fields, sliders, and menus. In general, these objects already

have meanings to most computer users. For example, when you move a slider, value changes; when you press an OK button, your settings are applied and the dialog box is dismissed. Of course, to leverage this built-in familiarity, you mustbe consistent in how you use the various GUI building components.

The process of implementing a GUI involves two basic tasks:

- Laying out the GUI components
- Programming the GUI components

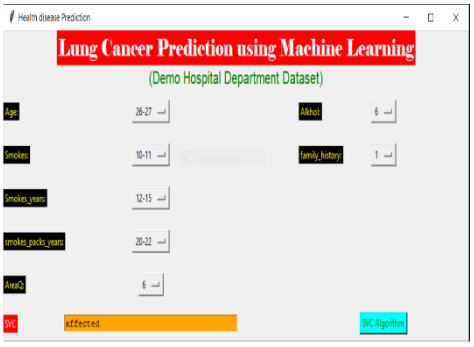
GUIDE primarily is a set of layout tools. However, GUIDE also generates an M-file that contains code to handle the initialization and launching of the GUI. This M-file provides a framework for the implementation of the call backs - the functions that execute when users activate components in the GUI

CONCLUSION

In this paper, we propose a machine learning-based method to accurately predict the lung cancer using supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface of lung cancer prediction by attributes.

RESULT

RESULT:



FUTURE WORK

To automate this process by show the prediction result in webapplication or desktop application.

To optimize the work to implement in Artificial Intelligence environment.

REFERENCE

[1] Torre L A, Siegel R L, Jemal A. Lung Cancer Statistics[J]. Advances in Experimental Medicine & Biology, 2015, 893:1-19.

[2] Chen W, Zheng R, Baade P D, et al. Cancer statistics in China, 2015[J]. CA: A Cancer Journal for Clinicians, 2016, 66(2):115-132.

[3] Chen W, Zheng R, Zeng H, et al. Epidemiology of lung cancer in C hina[J]. Thoracic cancer, 2015, 6(2): 209-215.

[4] None. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening[J]. New England Journal of Medicine, 2011, 365(5):395-409.

[5] de Carvalho Filho A O, Silva A C, de Paiva A C, et al. 3D shape analysis to reduce false positives for lung nodule detection systems[J]. Medical & biological engineering & computing, 2017, 55(8): 1199-1213.

[6] Yuan S, Ying W, Dazhe Z. Computer-Aided Lung Nodule Recognition by SVM Classifier Based on Combination of Random Under sampling and SMOTE[J]. Computational & Mathematical Methods in Medicine, 2015, 2015:1-13.

[7] Robert D. Ambrosini, Peng Wang, Walter G. O'Dell. Volume change determination of metastatic lung tumors in CT images using 3-D template matching[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2009, 7260.

[8] Ashwin S, Ramesh J, Kumar S A, et al. Efficient and reliable lung nodule detection using a neural networkbased computer aided diagnosis system[C]//2012 International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM). IEEE, 2012: 135-142.

[9] Tan M, Deklerck R, Jansen B, et al. A novel computer-aided lung nodule detection system for CT images[J]. Medical Physics, 2011, 38(10): 5630-5645.

[10] Lavanya K, Durai M A S, Iyengar N. Fuzzy rule based inference system for detection and diagnosis of lung cancer[J]. International Journal of Latest Trends in Computing, 2011, 2(1): 165-171.

[11] Kaya A, Can A B. A weighted rule based method for predicting malignancyof pulmonary nodules by nodule characteristics[J]. Journal of Biomedical Informatics, 2015, 56: 69-79.

[12] Kim B C, Yoon J S, Choi J S, et al. Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection[J]. Neural Networks, 2019, 115: 1-10.

[13] Lin-Lin W, Chun-Lei X. Hybrid method of image segmentation usingwatershed transform and improved FCM[J]. Computer Engineering & Applications, 2010, 46(14):189-191.

[14] Armato S G, Giger M L, Moran C, et al. Computerized Detection ofPulmonary Nodules on CT Scans[J]. Radiographics, 1999, 19(5):

[15] Okada K, Comaniciu D, Krishnan A. Robust anisotropic Gaussian fittingfor volumetric characterization of Pulmonary nodules in multislice CT[J]. IEEE Transactions on Medical Imaging, 2005, 24(3): p.409-423.

[16] Kostis W J, Reeves A P, Yankelevitz D F, et al. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images[J]. IEEE Transactions on Medical Imaging, 2003,2(10):1259-1274.

[17] Xie H, Yang D, Sun N, et al. Automated pulmonary nodule detection inCT images using deep convolutional neural networks[J]. Pattern Recognition, 2019, 85: 109-119.

[18] Ding J, Li A, Hu Z, et al. Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks[C]International Conference on Medical Image Computing and Computer-AssistedIntervention. Berlin: Springer, 2017:559-567.

[19] Dou Q, Chen H, Jin Y, et al. Automated Pulmonary Nodule Detection via 3D ConvNets with Online Sample Filtering and Hybrid-Loss ResidualLearning[C] International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2017: 630-638.

[20] Lv X Q, Wu L, Gu Y, et al. Detection of low dose CT pulmonary nodules based on 3D convolution neural network[J]. Optics and Precision Engineering, 2018, 26(5): 1211-1218.

[21] Zhao Y, Bock G H, Vliegenthart R, et al. Performance of computer-aided detection of pulmonary nodules in low-dose CT: comparison with double reading by nodule volume[J]. European Radiology, 2012, 22(10): 2076-2084.

[22] Zhang K, Zuo W, Chen Y, et al. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.

[23] Han H, Zhang H, Wei X, et al. Texture-preserved penalized weighted least- squares reconstruction of low-dose CT image via image segmentation and high- order MRF modeling[C]//Medical Imaging 2016: Physics of Medical Imaging. International Society for Optics and Photonics, 2016, 9783: 97834F.

[25] Chen,Geng,Xiang,Dehui ,Zhang,Bi,etal.AutomaticPathologicalLung Segmentation in Low-dose CT Image using Eigenspace Sparse Shape Composition[J]. IEEE Transactions on Medical Imaging:1-1.

[26] Tschirren, J, Hoffman, E.A, McLennan, G,et al. Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose CT scans[J]. IEEE Transactions on Medical Imaging, 24(12): p. 1529-1539.

[27] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C] //International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[28] Yang H, Yu H, Wang G. Deep learning for the classification of lung nodules[J]. arXiv preprint arXiv:1611.06651, 2016. ELCAP Public Lung Image Database:[DB/OL]:http://www.via.cornell. edu/databases/lungdb.html.

[29] Setio A A A, Ciompi F, Litjens G, et al. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1160-1169.

[30] Da Silva G L F, Silva A C, de Paiva A C, et al. Classification of Malignancyof Lung Nodules in CT Images Using Convolutional Neural Network[J]. Workshop de Informática Médica, 2016(16): 2481-2489.

[31] Kumar D, Wong A, Clausi D A. Lung Nodule Classification Using Deep Features in CT Images[C]//Computer and Robot Vision. IEEE, 2015:133-138.

[32] LIU Lu, LIU Wan-yu,CHU Chun-yu, WU Jun,ZHOU Yang,ZHANG Hong-xia,BAO Jie. Fast classification of benign and malignant solitary pulmonary nodules in CT image[J]. Optics and Precision Engineering, 2009, 17(8): 2060- 2068

[33] Orozco H M, Villegas O O V, Sánchez V G C, et al. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine[J]. Biomedical Engineering Online, 2015, 14(1):9.

[34] Farag A A, Elhabian S Y, Elshazly S A, et al. Quantification of nodule detection in chest CT: A clinical investigation based on the ELCAP study[C]//Proc. of Second International Workshop on Pulmonary Image Processing in conjunction with MICCAI. 2009: 149-160.

[35] Hosseini M P, Pompili D, Elisevich K, et al. Optimized deep learning for EEG big data and seizure prediction BCI via internet of things[J]. IEEE Transactions on Big Data, 2017, 3(4): 392-404.

[36] Puthal D. Lattice-modeled information flow control of big sensing data streams for smart health application[J]. IEEE Internet of Things, 2019, 6(2): 1312-1320.

[37] Pasluosta C F, Gassner H, Winkler J, et al. An emerging era in the management of Parkinson's disease: wearable technologies and the internet of

[38] things[J]. IEEE Journal of Biomedical and Health Informatics, 2015, 19(6): 1837-1881.

[39] Jara A J, Zamora-Izquierdo M A, Skarmeta A F. Interconnection framework for mHealth and remote monitoring based on the internet of things[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(9): 47-65.

[40] Chen M, Yang J, Zhou J, et al. 5G-smart diabetes: toward personalized diabetes diagnosis with healthcare big data clouds[J]. IEEE Communications Magazine, 2018, 56(4): 16-23.

[41] Hossain M S, Muhammad G. Emotion-aware connected healthcare big data towards 5G[J]. IEEE Internet of Things Journal, 2018, 5(4): 2399-2406.

[42] Kim J, Lee W. Stochastic decision making for adaptive crowdsourcing in medical big-data platforms[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(11): 1471-1476.

[43] Chowdhury M Z, Shahjalal M, Ahmed S, et al. 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions. DOI: arXiv:1909.11315