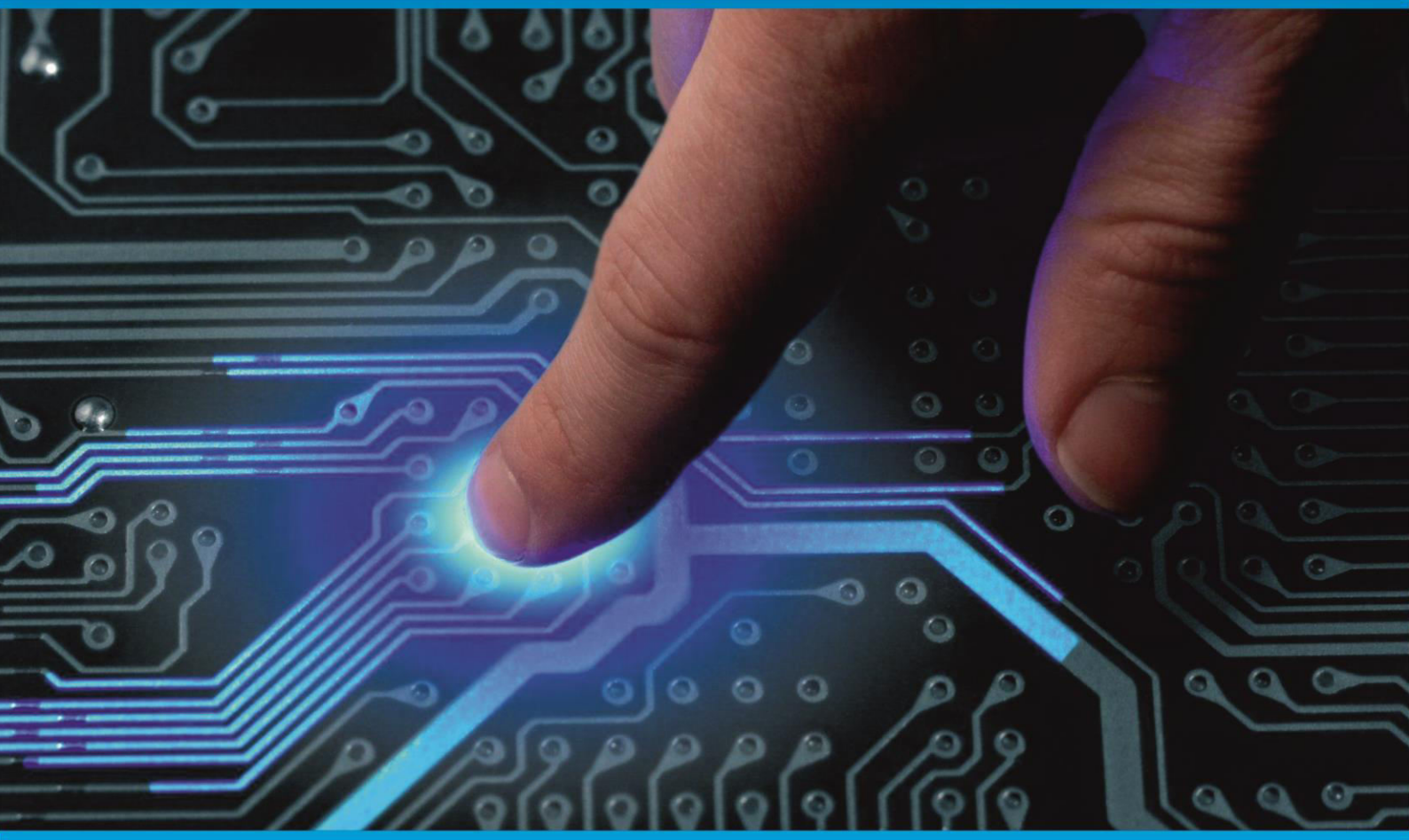
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 3, March 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

|e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

|DOI: 10.15680/IJIRCCE.2021.0903197|

Diagnose of Lung Cancer Prediction Using Ensemble Learning

Mr.V.Vinothkumar,M.Tech[1] Ashwinkumar.A [2], Ganesh.R [3], Krishnan.A [4]

Asst. Professor, Gojan School of Business and Technology, Edapalaiyam, Redhills, Tamilnadu, India

Gojan School of Business and Technology, Edapalaiyam, Redhills, Tamilnadu, India

Gojan School of Business and Technology, Edapalaiyam, Redhills, Tamilnadu, India

Gojan School of Business and Technology, Edapalaiyam, Redhills, Tamilnadu, India

ABSTRACT: 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. To 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.

KEYWORDS: lung cancer detection, lung cancer prediction,SVM,Machine Learning

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. 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 be done on the entire given dataset.To 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.

II. METHODOLOGY

A. Data validation process

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 error rate 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.

B. Data 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 dataset and 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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903197|

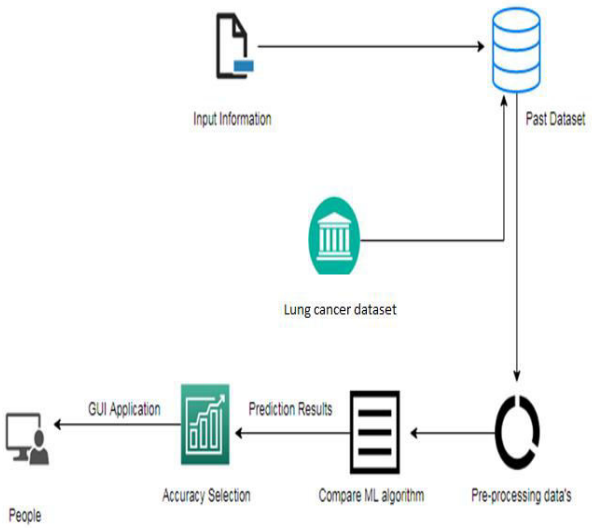
C. Comparison&Accuracy of algorithm

To train a model by given dataset using sklearn package comparison&accuracy result of algorithm.

D. GUI based on result

Tkinder is a python library for developing GUI (Graphical User Interface).we use the tkinder library for creating an application of UI(User Interface),to create windows and all other graphical user interface and tkinder will come with python as a standard package,it can be used for security purpose of each users or accountants.

III. SYSTEM ARCHITECTURE



IV. DEVELOPMENT TOOLS

This chapter is about the software language and the tools used in the development of the project. The platform used here is PYTHON.

A. The Python framework

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structure, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development. As well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduce the cost ofprogaram maintenance.Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary from without charge for all all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy. a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error. It raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on the debugger is written Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source the fast edit-test-debug cycle makes this simple.

B. Pandas in overflow

Pandas is software library written for the Python programming languages for manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data" an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the pharse "Python data analysis" itself. Wes McKinney started building what would because pandas at AQR Capital while he was a researcher there from 2007 to 2010.Data Frame object for data manipulation with integrated indexing. Tools for reading and writing data between in-memory data structures and different file formats. Data alignments and integrated

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903197|

handling of missing data. Reshaping and pivoting of data sets. Labe 1-based slicing, fancy indexing, and sub setting of large data sets. Data structure column insertion and deletion. Group by engine allowing split-apply-combine operations on data sets. Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure. Time series-functionally Data range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging Provides data filtration. The library is highly optimized for performance, with critical code paths written in Python or C. pandas is mainly used for data analysis. Pandas allows importing data from data from various file formats such as comma separated values, JSON, SQL, Microsoft Excel.

C. Software Description

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications,large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system "Conda". The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the `conda install` command or using the `pip install` command that is installed with Anaconda. Pip packages provide many of the features of conda packages and in most cases they can work together. Custom packages can be made using the `conda build` command, and can be shared with others by uploading them to Anaconda Cloud, PyPI or other repositories. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, you can create new environments that include any version of Python packaged with conda.

Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

D. Conda

Conda is an open source, cross-platform, language-agnostic package manager and environment management system that installs, runs and updates packages and their dependencies. It was created for Python programs, but it can package and distribute software for any language (e.g., R), including multi-languages. The Conda package and environment manager is included in all versions of Anaconda, Miniconda, and Anaconda Repository.

E. The Jupyter notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

F. Kernal

A notebook kernel is a "computational engine" that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels). When you open a Notebook document, the associated kernel is automatically launched. When the notebook is executed (either cell-by-cell or with menu Cell -> Run All), the kernel performs the computation and produces the results. Depending on the type of computations, the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel is shut-down.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903197|

G. Notebook Dashboard

The Notebook Dashboard is the component which is shown first when you launch Jupyter Notebook App. The Notebook Dashboard is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown). The Notebook Dashboard has other features similar to a file manager, namely navigating folders and renaming/deleting files.

H. Working process

- Download and install anaconda and get the most useful package for machine learning in Python.
- Load a dataset and understand its structure using statistical summaries and data visualization.
- machine learning models, pick the best and build confidence that the accuracy is reliable.

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming. The best way to get started using Python for machine learning is to complete a project.It will force you to install and start the Python interpreter (at the very least).It will give you a bird's eye view of how to step through a small project. It will give you confidence, maybe to go on to your own small projects. When you are applying machine learning to your own datasets, you are working on a project. A machine learning project may not be linear, but it has a number of well-known steps:

- Define Problem.
- Prepare Data.
- Evaluate Algorithms.
- Improve Results.
- Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely, from loading data, summarizing data, evaluating algorithms and making some predictions. Here is an overview of what we are going to cover:Installing the Python anaconda platform.

1. Loading the dataset.
2. Summarizing the dataset.
3. Visualizing the dataset.
4. Evaluating some algorithms.

V. SOFTWARE TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

A. Developing Methodologies

The test process is initiated by developing a comprehensive plan to test the general functionality and special features on a variety of platform combinations. Strict quality control procedures are used.The process verifies that the application meets the requirements specified in the system requirements document and is bug free. The following are the considerations used to develop the framework from developing the testing methodologies.

B. Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program input produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903197|

C. Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:Valid Input: identified classes of valid input must be accepted. Invalid Input: identified classes of invalid input must be rejected. Functions: identified functions must be exercised. Output: identified classes of application outputs must beexercised. Systems/Procedures: interfacing systems or procedures must be invoked.

D. System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

E. Performance Testing

The Performance test ensures that the output be produced within the time limits, and the time taken by the system for compiling, giving response to the users and request being send to the system for to retrieve the result.

F. Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

G. Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements. Acceptance testing for data synchronization:

H. Build the test plan

Any project can be divided into units that can be further performed for detailed processing. Then a testing strategy for each of this unit is carried out. Unit testing helps to identity the possible bugs in the individual component, so the component that has bugs can be identified and can be rectified from errors.

V. CONCLUSION

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. To 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. As the result, our system was capable to detect and identify whether the breath were from lung cancer patients or not.

REFERENCES

[1]Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971-987.

[2]Radüntz, T., Scouten, J., Hochmuth, O., & Meffert, B. (2015). EEG artifact elimination by extraction of ICA-component features using image processing algorithms. *Journal of neuroscience methods*, *243*, 84-93.

[3]Sairamya, N. J., Selvaraj, T. G., Ramasamy, B., Deivendran, N. P., & Subathra, M. S. P. (2018). Classification of EEG signals for detection of epileptic seizure activities based on feature extraction from brain maps using image processing algorithms. *IET Image Processing*.

[4] M. U. Dalmış, A. Gubern-Mérida, S. Vreemann, N. Karssemeijer, R.Mann, and B. Platel, "A computer-aided diagnosis system for breastDCE-MRI at high spatiotemporal resolution," *Medical physics,* vol. 43,pp. 84-94, 2016.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation,"*arXiv preprint arXiv:1511.00561,* 2015.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903197|

[6] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J.Kirby, *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging,* vol. 34, pp. 1993-2024, 2015.
[7] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 447-456, 2015.
[8] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2017," CA: A Cancer Journal for Clinicians, vol. 67, no. 1,pp. 7–30, 2017.
[9] Heber MacMahon, John HM Austin, Gordon Gamsu, Christian J Herold, James R Jett, David P Naidich, Edward F Patz Jr, and Stephen J Swensen, "Guidelines for management of small pulmonary nodules detected on ct scans: a statement from the fleischner society," Radiology, vol. 237, no. 2, pp. 395–400, 2005.
[10]Sairamya, N. J., Selvaraj, T. G., Ramasamy, B., Deivendran, N. P., & Subathra, M. S. P. (2018). Classification of EEG signals for detection of epileptic seizure activities based on feature extraction from brain maps using image processing algorithms. IET Image Processing.
[11] Sairamya, N. J., George, S. T., Subathra, M. S. P., & Kumar, N. M. (2019). Computer-Aided Diagnosis of Epilepsy Based on the Time Frequency Texture Descriptors of EEG Signals Using Wavelet Packet Decomposition and Artificial Neural Network. In Cognitive Informatics and Soft Computing (pp. 677-688). Springer, Singapore.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details