# PREDICTION OF LUNG CANCER LEVELS BASED ON PATIENT LIFESTYLE AND HISTOPATHOLOGICAL IMAGES USING ARTIFICIAL INTELLIGENCE

**SOFYAN EL IDRISSI[1], IKRAM BEN ABDEL OUAHAB[2], YASSINE DRIDER[3], MOHAMMED BOUHORMA[4], FATIHA EL OUAAI[5]**

Computer science, systems and telecommunication laboratory (LIST),

University Abdelmalek Essaadi, FSTT, Tangier, Morocco

E-mail: [1]idrissisofyan38@gmail.com, [2]ibenabdelouahab@uae.ac.ma, [3]yassinedrider@gmail.com
[4]mbouhorma@uae.ac.ma, [5]felouaai@uae.ac.ma

## ABSTRACT

Histopathology is the fundamental tool used in pathology for more than a century to establish the final diagnosis of bronchopulmonary carcinoma. The phenotypic information on histological images reflects the overall effect of molecular alterations on the behavior of cancer cells and provides a convenient visual readout of disease aggressiveness. However, the human assessment of the histological image can be sometimes subjective and not very reproducible depending on the case. Therefore, computational analysis of histological imaging via artificial intelligence approaches has recently received considerable attention to improve this diagnostic accuracy. Thus, computational analysis of lung cancer images has recently been evaluated for optimization of histological or cytological classification, prognosis prediction or genomic profiling of lung cancer patients. This rapidly growing field is constantly showing great power in the field of medical imaging informatics by producing detection, segmentation or recognition tasks of a very high accuracy. However, there are still several major challenges or issues to be addressed in order to successfully transfer this new approach into clinical routine. In this paper, we use the power of AI to develop an GUI application able to predict lung cancer levels using two techniques: either clinical features based on the patient lifestyle, or by inserting histopathological lung images. The GUI application is friendly to use, fast and average accuracy of 98%. We also did a comparative study of machine learning and deep learning algorithms for lung cancer classification using 2 different databases, then we choose the best ones to be used.

**Keywords:** *Pathology, Histology, Cytology, Bronchopulmonary Cancer, Artificial Intelligence, Machine Learning.*

## 1. INTRODUCTION

Lung cancer, also called bronchial cancer or bronchopulmonary cancer, is a disease of the cells of the bronchial tubes or, more rarely, of the cells that line the alveoli of the lungs. It develops from an initially normal cell that changes and multiplies in an uncontrolled way, until it forms a mass called a malignant tumor. Lung cancer develops from cells in the bronchi. There are two main types of lung cancer depending on the origin of the bronchial cells from which they arise: non-small cell lung cancer (NSCLC), which accounts for about 85% of lung cancers; and small cell lung cancer (SCLC), which accounts for about 15% of lung cancers.

Cancer is one of the leading causes of death in every country in the world and in 2020, lung cancer remained the leading cause of cancer deaths, with an estimated 1.8 million deaths (18% of all cancer deaths), according to the Global Cancer Statistics 2020 report from the American Cancer Society (ACS) and the International Agency for Research on Cancer (IARC). The study examines the global cancer burden in 2020 based on GLOBOCAN estimates of cancer incidence and mortality from the International Agency for Research on Cancer in 185 countries worldwide. In 2020, female breast cancer was the most frequently diagnosed cancer, with about 2.3 million new cases (11.7%), followed by lung (11.4%), colorectal (10%), prostate (7.3%), and stomach (5.6%) cancers. Over the past two decades, the total number of people diagnosed with cancer has nearly doubled from about 10 million in 2000 to 19.3 million in 2020. Today, one in five people worldwide will develop cancer in their lifetime.

Lung cancer is the 3rd most common cancer and the 1st cause of cancer death. Smoking is the main risk factor for lung cancer: a smoker is 10 to 15 times more likely to develop lung cancer than a non-smoker. Passive smoking increases a non-smoker's risk of developing lung cancer by 26%. The 2nd cause of lung cancer is an environmental factor: radon, responsible for about 3000 deaths per year. Approximately 15% of lung cancers have an occupational origin, the main occupational factor being asbestos. Lung cancer can also manifest itself by general symptoms: fatigue, weight loss, loss of appetite, prolonged fever, headaches, phlebitis, nervous disorders with confusion, progressive swelling of the tips of the fingers like "drumsticks". Since these symptoms are not typical of cancer, it is important to talk to your doctor, especially if they persist for several days or if they appear in a smoker. For non-small cell lung cancer, there are 5 stages: stage 0 followed by stages 1 to 4. For stages 1 to 4, the Roman numerals I, II, III and IV. In general, the higher the number, the more the cancer has spread.

The selection of biological data contributes towards the reinforcement of the medical diagnosis aid, the level and the rate of progression of biomarkers measured in a repetitive way on each subject allowing to quantify the severity of the disease and the susceptibility of its progression; this is usually interesting, on the clinical and scientific levels, to help the expert to take these decisions in a less late time than the survival of a patient. The significant increase in clinical and biological endpoints to be included in a therapeutic trial in the era of precision medicine makes dedicated trials impossible. Conducting phase 3 clinical trials in radiotherapy, comparing different treatment regimens or techniques (typically a conformal or intensity modulated technique treatment) may be difficult or even impossible in some cases, either because the older technique is no longer available or because it would be unethical to propose a randomized trial comparing these techniques. Finally, there is a significant difference between the patient populations included in clinical trials and the patients encountered in clinical practice.

New approaches are therefore needed in order to generate studies with a good level of evidence, and one such approach is the construction and exploitation of detailed massive databases, integrating a large number of parameters[1]. Within these data, molecular biology and genomics are set to play a prominent role. Prediction of radiosensitivity of healthy tissues and tumors will need to integrate the entire genome analysis of patients[2]. The complexity of the predictive models obtained will make it impossible to simply analysis by a human, as human cognitive integration capabilities are considered to take into account up to five factors, but by 2020, a medical decision may be based on 10,000 parameters at once for a single patient [3]. While the price of genetic sequencing has continuously decreased in parallel with the price of computing power, the only factor preventing us from generating these precise predictive models is the existence of large cohorts of patients whose precise phenotypic profile is known.

On the other hand, Deep Learning methods are widely applied to various fields of science and engineering such as speech recognition, image classification and learning methods in language processing. Similarly, traditional data processing techniques have several limitations in processing large amounts of data. In addition, Big Data analysis requires new and sophisticated algorithms based on Machine Learning and Deep Learning techniques to process data in real time with high accuracy and efficiency. However, recently, research has incorporated various Deep Learning techniques with different learning and training mechanisms for high-speed data processing. Deep Learning is a form of Artificial Intelligence, derived from Machine Learning, which uses many layers of artificial neurons that interact to allow computers to learn progressively and efficiently from various types of data.

In this paper, we propose 2 AI applications to deal with lung cancer prediction. First application using machine learning to predict lung cancer levels based on the patients' lifestyle, and we study also the factors that influence more the cancer level. Second application use deep learning to classifier lung cancer based on histopathological given images. Finally, to keep it simple to use, we developed a Graphical User Interface application, which has the form of a form to fill in and which will provide the results obtained by each classifier. The proposed application is fast and accurate, and will save time, effort and lives.

## 2. RELATED WORKS

In this section, we present the state of the art while summarizing the research work done in the field of oncology and the application of artificial intelligence in medicine.

In this paper [4], we find the early diagnosis of lung cancer by examining the performance of classification algorithms classification algorithms such as Naive Bayes, SVM, decision tree and logistic regression.

On the other hand [5], the author used a pioneering interdisciplinary mechanism, which is applied to lung cancer for the first time, to detect early diagnostic biomarkers of lung cancer by combining metabolomics and machine learning methods such as Naïve Bayes.

Regarding the article [6],the author asserts that the leading cause of cancer-related mortality in the world is "lung cancer". Therefore, the prior detection, prediction and diagnosis of lung cancer have become essential as they accelerate and simplify the resulting clinical picture. To erect the progress and medication of cancer diseases, machine learning techniques have been used due to their accurate results. Different types of machine learning (ML) algorithms such as Naive Bayes, Support Vector Machine (SVM), logistic regression, artificial neural network (ANN) have been applied in the health sector for the analysis and prognoses of lung cancer.

Thus, the paper [7] , it is noted that the nature of clinical data makes it difficult to quickly select, tune and apply machine learning algorithms to clinical prognosis. As a result, much time is spent searching for the most appropriate machine learning algorithms applicable to clinical prognosis that contain binary or multi-valued attributes. The purpose of this study was to identify and evaluate the performance of Machine learning classification applied to clinical prognosis of postoperative life expectancy in lung cancer patients. Multilayer Perceptron, J48, and Naive Bayes algorithms were used to train and test models on thoracic surgery datasets obtained from the University of California, Irvine machine learning repository. A 10-fold stratified cross-validation was used to evaluate the accuracy of the classifiers' baseline performance. The comparative analysis shows that Multilayer Perceptron performed the best with a classification accuracy of 82.3%, J48 came second with a classification accuracy of 81.8%, and Naive Bayes was the worst with a classification accuracy of 74.4%. The quality and outcome of the chosen machine learning algorithms depends on the ingenuity of the clinical miner.

Compared to the article [8] Early diagnosis has been shown to improve survival rates of lung cancer patients. The availability of blood-based screening could increase the early involvement of lung cancer patients. Our present study attempted to discover plasma metabolites of Chinese patients as diagnostic biomarkers of lung cancer. In this work, we use a pioneering interdisciplinary mechanism, which is first applied to lung cancer, to detect early diagnostic biomarkers of lung cancer by combining metabolomics and machine learning methods.

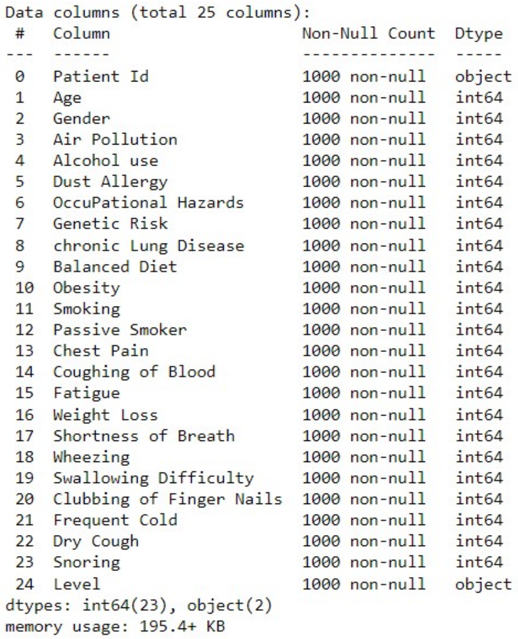
The paper [9] empirically evaluates several adaptable machine learning algorithms for lung cancer detection related to IoT devices. In this work, a review of nearly 65 papers for predicting different diseases, using machine learning algorithms, was conducted.

The analysis mainly focuses on various machine learning algorithms used to detect several diseases to look for a gap towards future improvement in lung cancer detection in medical IoT. Each technique has been analyzed at each stage, and the overall drawbacks are highlighted. In addition, it also analyzes the type of data used to predict the relevant disease, whether it is reference data or manually collected data. Finally, research directions have been identified and developed from the different existing methodologies.

## 3. METHODS

### 3.1 Lung Cancer Levels database

First, we will use a publicly available digital database on Kaggle [10] to make prediction of the level of cancer in the patient based on their lifestyles (clinical data such as: age, gender, alcohol consumption, allergy, weight, smoking, pain, and others).



```
Data columns (total 25 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Patient Id                1000 non-null   object
 1   Age                       1000 non-null   int64
 2   Gender                    1000 non-null   int64
 3   Air Pollution             1000 non-null   int64
 4   Alcohol use               1000 non-null   int64
 5   Dust Allergy              1000 non-null   int64
 6   OccuPational Hazards      1000 non-null   int64
 7   Genetic Risk              1000 non-null   int64
 8   chronic Lung Disease      1000 non-null   int64
 9   Balanced Diet             1000 non-null   int64
 10  Obesity                   1000 non-null   int64
 11  Smoking                   1000 non-null   int64
 12  Passive Smoker            1000 non-null   int64
 13  Chest Pain                1000 non-null   int64
 14  Coughing of Blood         1000 non-null   int64
 15  Fatigue                   1000 non-null   int64
 16  Weight Loss               1000 non-null   int64
 17  Shortness of Breath       1000 non-null   int64
 18  Wheezing                  1000 non-null   int64
 19  Swallowing Difficulty     1000 non-null   int64
 20  Clubbing of Finger Nails  1000 non-null   int64
 21  Frequent Cold             1000 non-null   int64
 22  Dry Cough                 1000 non-null   int64
 23  Snoring                   1000 non-null   int64
 24  Level                     1000 non-null   object
dtypes: int64(23), object(2)
memory usage: 195.4+ KB
```

*Figure 1: Lung cancer levels database structure*

To see the distribution of the data on the different classes, the bar chart for the number of patients for each cancer level. The data are almost balanced, since we find approximately the same values for each class. In the x-axis (Cancer levels): Low, Medium, High. On the y-axis: the total number of patients for each level.

Next, we wanted to see the distribution of the data by gender of the patients. The bar chart for the number of patients for each level of cancer and for each gender (male and female). The 1 represents the men (in blue). The 2 represents women (in red). It is noticeable that in the present database, a high number of men suffer from high-grade lung cancer. It can be assumed that this is due to the fact that men smoke more than women.

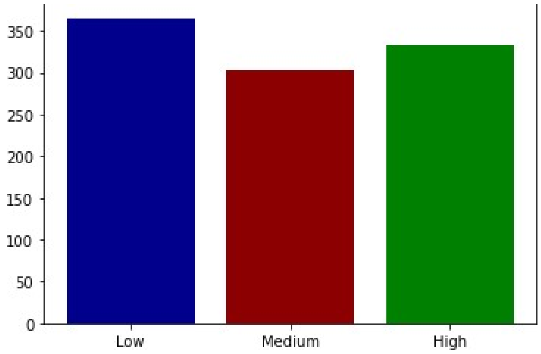


*Figure 2: lung cancer levels : data distribution*

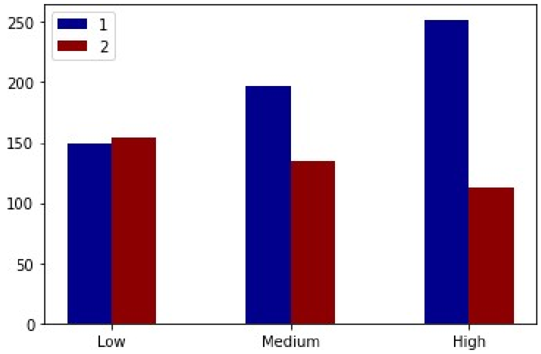


*Figure 3: Distribution of data by gender (male/female)*

### 3.2 Histopathological images of lung cancer database

A database was used that contains publicly available images on Kaggle [11]. This dataset contains 25,000 histopathology images with 5 classes. All images are 768 x 768 pixels in size and are in jpeg file format. Images were generated from an original sample of HIPAA compliant and validated sources, consisting of 750 total lung tissue images (250 benign lung tissue, 250 lung adenocarcinoma, and 250 lung squamous cell carcinoma) and 500 total colon tissue images (250 benign colon tissue and 250 colon adenocarcinoma) and augmented to 25,000 using augmentation. There are five classes in the dataset, each with 5,000 images, namely:

- Benign lung tissue
- Lung adenocarcinoma
- Pulmonary squamous cell carcinoma
- Adenocarcinoma of the colon
- Benign tissue of the colon

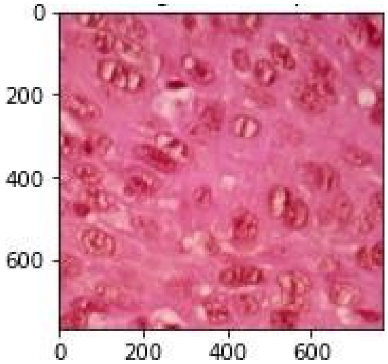


*Figure 4 Lung SCC sample*

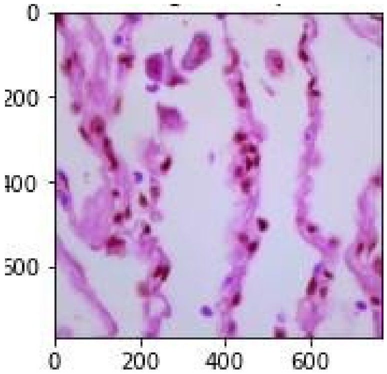


*Figure 5 Lung N sample*

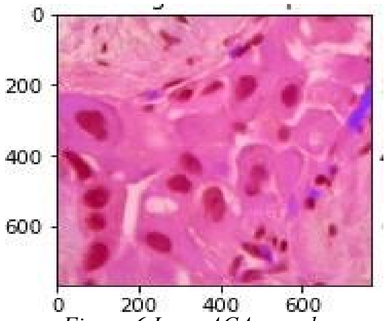


*Figure 6 Lung ACA sample*

### 3.3 Machine Learning for Lung Cancer Level Prediction

In order to predict the lung cancer level, we train 3 machine learning models using the numerical features given in the database. These models are :

- K-Nearest Neighbor,
- Decision Tree,
- Random Forest.

*Table 1: Machine learning algorithms used to prediction lung cancer levels*

| ***Algorithm*** | ***Definition*** |
|---|---|
| **KNN** | K-nearest neighbors algorithm is a Machine Learning algorithm that belongs to the class of simple and easy to implement supervised learning algorithms that can be used to solve classification and regression problems. In this article, we will review the definition of this algorithm, its operation and a direct application in programming. In supervised learning, an algorithm receives a set of data that is labeled with corresponding output values on which it can train and define a prediction model. This algorithm can then be used on new data to predict their corresponding output values. |
| **DT** | Decision tree is a supervised learning technique that can be used for both classification and regression problems, but it is generally preferred for solving classification problems. It is a tree-based classifier, where the internal nodes represent the features of a data set, the branches represent the decision rules, and each leaf node represents the result. In a decision tree, there are two nodes, which are the decision node and the leaf node. Decision nodes are used to make any decision and have multiple branches, while leaf nodes are the output of those decisions and contain no other branches. The decisions or test are made on the basis of the characteristics of the given data set. It is a graphical representation of all possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which grows into other branches and builds a tree structure.<br>To build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question and, based on the answer (Yes/No), it then divides the tree into subtrees. |
| **RF** | Random Forest is a very popular Machine Learning technique among Data Scientists and for good reason: it has many advantages compared to other data algorithms. It is an easy-to-interpret, stable technique, which generally has good accuracies and can be used for regression or classification tasks. It covers a large part of the Machine Learning problems. In Random Forest there is first the word "Forest". So, this algorithm will be based on trees that we call decision tree or decision tree. As the name suggests, a decision tree helps the data scientist make a decision through a series of questions (also called tests) whose answer (yes/no) will lead to the final decision. |
| **LR** | Logistic regression is a statistical model for studying the relationships between a set of qualitative variables Xi and a qualitative variable Y. It is a generalized linear model using a logistic function as a link function. A logistic regression model can also predict the probability of an event occurring (value of 1) or not (value of 0) from the optimization of the regression coefficients. This result always varies between 0 and 1. When the predicted value is above a threshold, the event is likely to occur, while when this value is below the same threshold, it is not. |

| | |
|---|---|
| **GNB** | Naïve Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data. |

### 3.4 Deep Learning for Histopathological images classification

Over the past decades, Deep Learning has proven to be a very powerful tool due to its ability to handle large amounts of data. The interest in using hidden layers has surpassed traditional techniques, especially in pattern recognition. One of the most popular deep neural networks is convolutional neural networks. Convolutional Neural Networks were first developed and used in the 1980s. The most a CNN could do at that time was to recognize handwritten numbers. It was mainly used in the postal sector to read postal codes, PIN codes, etc. The important thing to remember about any deep learning model is that it requires a large amount of data to train and also requires a lot of computing resources. This was a major drawback for CNNs at the time and as a result, CNNs were only limited to the postal sectors and failed to enter the machine learning world.

Transfer Learning is a method of machine learning where a model developed for one task is reused as the starting point for a model for a second task. This is a popular approach in the field of Deep Learning, where pre-trained models are used as a starting point for computer vision and natural language processing tasks, given the vast time and computational resources required to develop neural network models on these problems and the huge leaps in skill they provide on related problems. Moreover, ResNet101 [12] is a convolutional neural network with a depth of 101 layers. You can load a pre-trained version of the network, trained on over a million images from the ImageNet database. The pre-trained network can classify images into 1000 object categories, such as keyboard, mouse, pencil and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224 by 224. The central idea of ResNet is the introduction of a so-called "identity shortcut connection" that allows one or more layers to be skipped developed by Microsoft Research team [13].

In this paper, we use 4 pre-trained models : MobileNetV2, ResNet101, VGG16, and DenseNet169. We implement these models for transfer learning using histopathological lung cancer images. Which gives very good results.

## 4. PROPOSED SOLUTION

Our main contribution is to create a desktop application that takes a patient's information to predict the level of Lung Cancer based on machine learning. Also, we take a histopathology image to predict if our tumor is benign or malignant, if it is malignant, the application specifies the type of malignancy. Firstly, the user has the possibility to choose his prediction method, either the prediction of the Cancer level from the patient's information, or the prediction from a histopathological image.

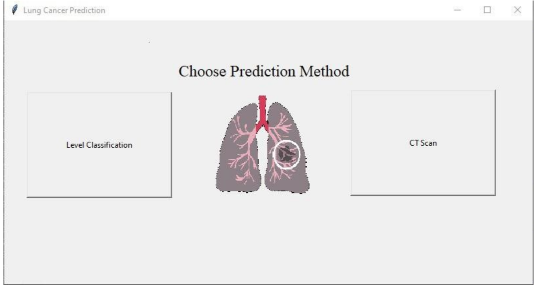
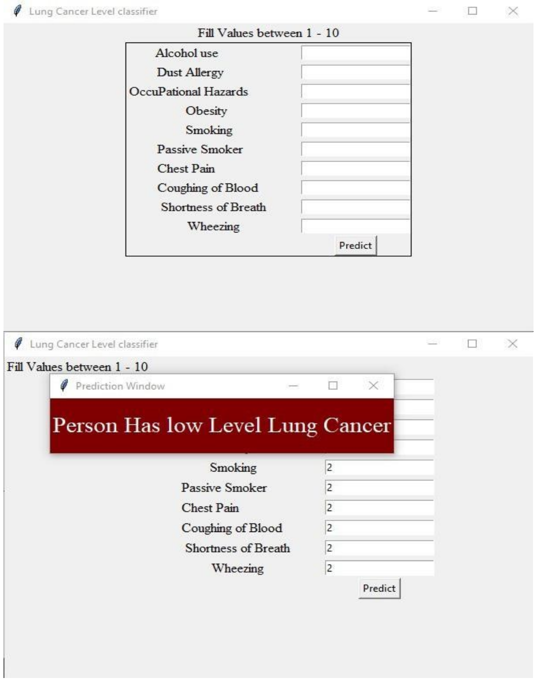


*Figure 7 Lung cancer prediction application*

If the user selects the Level Classification option, he will be directed to a new window, where he is asked to enter the patient's information. Noting that the requested information are the important Features selected in the Feature Selection part. After the user enters the patient's information and clicks on the 'Predict' button, a window appears with the prediction result.

- *"Person has **Low** Level Lung Cancer": If the patient has a low level of cancer*
- *"Person has **Medium** Level Lung Cancer": If the patient has a medium level of cancer*
- *"Person has **High** Level Lung Cancer": If the patient has a high level of cancer*



Fill Values between 1 - 10

Alcohol use
Dust Allergy
OccuPational Hazards
Obesity
Smoking
Passive Smoker
Chest Pain
Coughing of Blood
Shortness of Breath
Wheezing

Fill Values between 1 - 10

Person Has low Level Lung Cancer

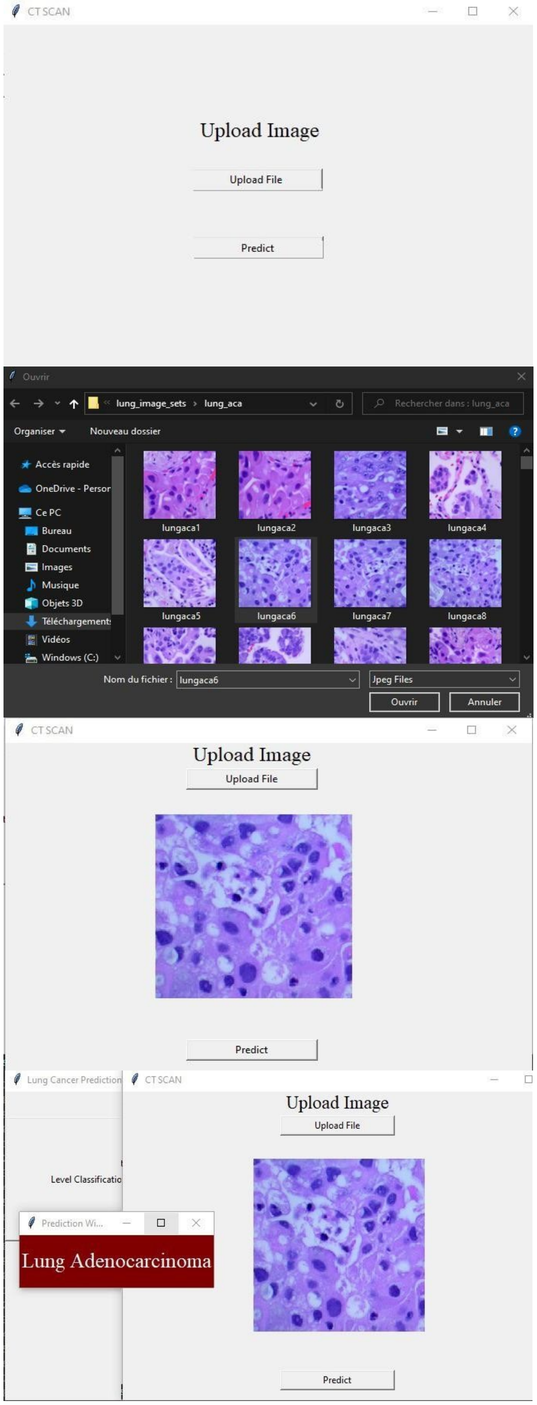
Smoking 2
Passive Smoker 2
Chest Pain 2
Coughing of Blood 2
Shortness of Breath 2
Wheezing 2

*Figure 7 Level classification of lung cancer*

If the user has chosen the classification of histopathological images and clicks on 'CT Scan', he will be directed to a new window. In this window, the user has the possibility to choose a histopathological image. After selecting the histopathological image, the desired image will be displayed in the window. By clicking on the 'Predict' button, a window containing the result will appear. The result is one of these 5 classes:

- *Colon Adenocarcinomas*
- *Benign Colonic Tissues*
- *Lung Adenocarcinomas*
- *Lung Squamous Cell Carcinomas*
- *Benign Lung Tissues*

This is the desktop application we developed, it takes data from patients to predict the level of Lung Cancer they have, and also take a histopathological image to predict if our tumor is benign or malignant, if it is malignant, the application specifies the type of malignancy.



Nom du fichier : lungaca6
Jpeg Files

Lung Adenocarcinoma

*Figure 8 CT Scan Classification Application*

## 5. RESULTS AND DISCUSSION

To get the best lung cancer level prediction model, we evaluated 5 machine learning algorithms: GNB, DT, RF, LR, and KNN. Then, we evaluated these models using evaluation metrics as the confusion matrix. The results of this task is given in table 2 and Figure 10. We use grid search to find the appropriate hyperparameter of each algorithm. Then, the worst classifier is GNB with 92% test accuracy. The best result goes to LR with 99% training score and 98% testing accuracy. Regarding DT and RF we suppose that it could be overfitted somewhere. As a result, we consider the LR as our best accurate classifier that we'll be using in the GUI application.

*Table 2 Comparison of machine learning algorithms for lung cancer level prediction task*

| ***Model*** | ***Train_Score*** | ***Test_accuracy*** |
|---|---|---|
| GNB | 0.880000 | 0.924 |
| DT | 1.000000 | 1.000 |
| RF | 1.000000 | 1.000 |
| **LR** | **0.997333** | **0.984** |
| KNN | 0.969333 | 0.960 |

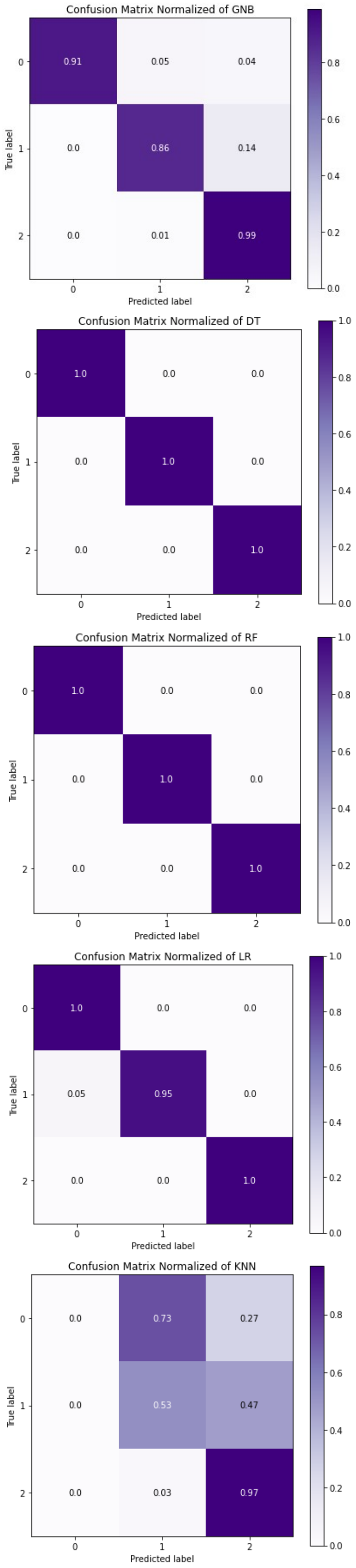


*Figure 9 Confusion matrix of lung cancer levels prediction using various machine learning models*

Moreover, we apply feature selection techniques on our database. We want to keep only the 10 most important features. Alcohol use, , Dust Allergy, Smoking, Balanced Diet, Obesity, Passive Smoker, Chest Pain, Coughing of Blood are important in the classification of lung cancer levels. Therefore, these are features that should be taken into consideration. However, these features: Air Pollution, OccuPational Hazards , Genetic Risk, fatigue Shortness of breath have a less important influence on the classification as they have appeared in some of the feature selection algorithms. The other features have no influence on the classification since they did not appear in any of the feature selection algorithms.

*Table 3 Results of the most 10 important features for lung cancer level prediction*

| ***Univariate Feature Selection (UFS)*** | ***Correlation matrix*** | ***PCA*** |
|---|---|---|
| Air Pollution | Air Pollution | Age |
| Alcohol use | Alcohol use | Alcohol use |
| Dust Allergy | Dust Allergy | Shortness of breath |
| OccuPational Hazards | OccuPational Hazards | OccuPational Hazards |
| Genetic Risk | Genetic Risk | Wheezing |
| Balanced Diet | Balanced Diet | Dry Cough |
| Obesity | Obesity | Air Pollution |
| Passive Smoker | Passive Smoker | Dry Cough |
| Chest Pain | Chest Pain | Smoking |
| Coughing of Blood | Coughing of blood | Swallowing Difficulty |

Now, regarding histopathological images of lung cancer classification, we evaluate the 4 pre-trained models using transfer learning on new data. The obtain results are given in table 4. In general, all models performed well. However, the best performance are shown while doing transfer learning to the VGG16 pretrained model. The test accuracy is the highest with 97.46%. So, we'll be using this model to predict new images in our GUI application. After training and evaluating the models, we saved the best one for further use.

*Table 4 Results of transfer learning use for histopathological lung cancer images classification*

| ***Model*** | ***Train acc.*** | ***Test acc.*** | ***F1-score*** | ***Recall*** | ***Precision*** |
|---|---|---|---|---|---|
| **MobileNet V2** | 92,81 | 89,82 | 89,74 | 89,82 | 89,81 |
| **ResNet 101** | 98,94 | 97,33 | 97,33 | 97,33 | 97,40 |
| **VGG 16** | **99,67** | **97,46** | **97,46** | **97,46** | **97,48** |
| **DenseNet 169** | 94,54 | 93,64 | 93,62 | 93,64 | 93,79 |

## 6. CONCLUSION

The computer systems that govern medical practice generate highly accurate patient data. This heterogeneous data includes demographic, socioeconomic, clinical, biological, imaging, and genomic information. The use of this data requires automated processes for standardization and integration into clinical data warehouses. The data specific to radiotherapy are very rich and of good quality. They are mainly derived from the mapping of the radiation dose received by the tumor and the surrounding organs. It also comes from prospective registration systems of treatments performed. These data are rarely integrated into clinical data and are therefore not used in predictive models.

We have therefore developed an application to extract the level of cancer in patients. The use of machine learning makes it possible to exploit this medical data in order to make a diagnosis or a prediction. In our proof-of-concept study, we created a predictive model using Deep Learning and Machine Learning on a cohort of patients with locally advanced lung cancer. This model can identify the level of cancer in this patient, as well as factors and symptoms, without the need for x-rays and scans. This type of approach could be used to personalize treatments and reduce their sequelae. However, validation of these algorithms will be an important and difficult step before they can be used in clinical routine. The work we present in this manuscript consists in the application of Machine Learning in the prediction of cancer levels based on the patients' lifestyle. We also study the factors that most influence the level of cancer level of these patients.

## REFERENCES

[1] T. Skripcak et al., « Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets », Radiother. Oncol., vol. 113, no 3, p. 303-309, déc. 2014, doi: 10.1016/j.radonc.2014.10.001.

[2] W. G. Jiang et al., « Tissue invasion and metastasis: Molecular, biological and clinical perspectives », Semin. Cancer Biol., vol. 35, p. S244-S275, déc. 2015, doi: 10.1016/j.semcancer.2015.03.008.

[3] A. P. Abernethy et al., « Rapid-Learning System for Cancer Care », J. Clin. Oncol., vol. 28, no 27, p. 4268-4274, sept. 2010, doi: 10.1200/JCO.2010.28.5478.

[4] R. P.R., R. A. S. Nair, et V. G., « A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms », in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), févr. 2019, p. 1-4. doi: 10.1109/ICECCT.2019.8869001.

[5] Y. Xie et al., « Early lung cancer diagnostic biomarker discovery by machine learning methods », Transl. Oncol., vol. 14, no 1, p. 100907, janv. 2021, doi: 10.1016/j.tranon.2020.100907.

[6] L. S.k., S. N. Mohanty, S. K., A. N., et G. Ramirez, « Optimal deep learning model for classification of lung cancer on CT images », Future Gener. Comput. Syst., vol. 92, p. 374-382, mars 2019, doi: 10.1016/j.future.2018.10.009.

[7] S. Kakeda et al., « Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System », AJR Am. J. Roentgenol., vol. 182, p. 505-10, mars 2004, doi: 10.2214/ajr.182.2.1820505.

[8] D. S. Ettinger et al., « Non-Small Cell Lung Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology », J. Natl. Compr. Cancer Netw. JNCCN, vol. 15, no 4, p. 504-535, avr. 2017, doi: 10.6004/jnccn.2017.0050.

[9] A. El-Baz et al., « Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies », Int. J. Biomed. Imaging, vol. 2013, p. 942353, janv. 2013, doi: 10.1155/2013/942353.

[10] Cancer Patients Data. Consulté le: 29 août 2022. [En ligne]. Disponible sur: https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

[11] Lung and Colon Cancer Histopathological Images. Consulté le: 29 août 2022. [En ligne]. Disponible sur: https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

[12] K. He, X. Zhang, S. Ren, et J. Sun, « Deep Residual Learning for Image Recognition ». arXiv, 10 décembre 2015. Consulté le: 9 septembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1512.03385

[13] K. He, X. Zhang, S. Ren, et J. Sun, « Identity Mappings in Deep Residual Networks ». arXiv, 25 juillet 2016. Consulté le: 9 septembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1603.05027