LUNG CANCER PREDICTION USING ML

A Project Report

Submitted in partial fulfilment of the

Requirements for the award of the Degree of

MASTER OF SCIENCE (COMPUTER SCIENCE)

By

SHAIKH SALMA MANJAR

Roll Number - 276

Under the esteemed guidance of

HOD Jasmeet Kaur Ghai

Assistant Prof. Rashmi Pote

Assistant Prof. Sunita Rai

DEPARTMENT OF COMPUTER SCIENCE

GURU NANAK KHALSA COLLEGE

OF

ARTS, SCIENCE & COMMERCE

(Autonomous)

MATUNGA, MUMBAI – 400 019

MAHARASHTRA

AY 2023 – 2024

GURU NANAK KHALSA COLLEGE
OF
ARTS, SCIENCE & COMMERCE
(Autonomous)
MATUNGA, MUMBAI, MAHARASHTRA – 400 019

DEPARTMENT OF COMPUTER SCIENCE

CERTIFICATE

This is to certify that the project entitled, " **Lung Cancer Prediction Using ML**", is bonafide topic of **Shaikh Salma Manjar** bearing Seat No: **276** submitted in partial fulfilment of MASTER OF SCIENCE (COMPUTER SCIENCE) Semester III

Topic Accepted ☐ Topic Rejected ☐ Need to be Discussed ☐ To be Discussed by External ☐

Signature of Student

Date: _____

HOD Computer Science

Date: _____

Signature of Prof. Rashmi Pote

Date: _____

Signature of External Examiner

Date: _____

Signature of Prof. Sunita Rai

Date: _____

Submitted by:

NAME : Shaikh Salma Manjar

ROLL NO : 276

CLASS : MSc. Comp Sci – Part-2

TOPIC : Lung Cancer Prediction Using ML

1.INTRODUCTION:

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets.

This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction.

Machine learning methods can be used to substantially (15-25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a more fundamental level, it is also evident that machine learning is also helping to improve our basic understanding of cancer development and progression.

Machine Learning in Cancer Research:

Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years. Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images to the classification of malignancies from proteomic and genomic.

Machine learning has been used primarily as an aid to cancer diagnosis and detection. It has only been relatively recently that cancer researchers have attempted to apply machine learning towards cancer prediction and prognosis.

Goal of Cancer prediction:

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci:

- ✓ The prediction of cancer susceptibility (i.e. risk assessment);
- ✓ The prediction of cancer recurrence and
- ✓ The prediction of cancer survivability

Indeed, a cancer prognosis typically involves multiple physicians from different specialties using different subsets of biomarkers and multiple clinical factors, including the age and general health of the patient, the location and type of cancer, as well as the grade and size of the tumour.

Types of Data needed for cancer Prediction:
Typically, histological (cell-based), clinical (patient-based) and demographic (population-based) information must all be carefully integrated by the attending physician to come up with a reasonable prognosis. Even for the most skilled clinician, this is not easy to do. Similar challenges also exist for both physicians and patients alike when it comes to the issues of cancer prevention and cancer susceptibility prediction.

Family history, age, diet, weight (obesity), high-risk habits (smoking, heavy drinking), and exposure to environmental carcinogens (UV radiation, radon, asbestos, PCBs) all play a role in predicting an individual's risk for developing cancer.

This information does not provide enough information to make robust predictions or prognoses. Ideally what is needed is some very specific molecular details about either the tumour or the patient's own genetic make-up.

With the rapid development of genomic (DNA sequencing, microarrays), proteomic (protein chips, tissue arrays, immuno-histology) and imaging (fMRI, PET, micro-CT) technologies, this kind of molecular-scale information about patients or tumours can now be readily acquired.

If these molecular patterns are combined with macro-scale clinical data (tumor type, hereditary aspects, risk factors), the robustness and accuracy of cancer prognoses and predictions improves even more.

Benefits of using Machine Learning in Cancer Prediction:
The use of computers (and machine learning) in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine.

This movement towards predictive medicine is important, not only for patients (in terms of lifestyle and quality-of-life decisions) but also for physicians (in making treatment decisions) as well as health economists and policy planners (in implementing large scale cancer prevention or cancer treatment policies)

Research Paper-1

Complete citation:	**Application of Machine Learning in Cancer prediction and Prognosis.** Joseph A. Cruz, David S. Wishart Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada T6G 2E8 (2006)
Key words:	Cancer, machine learning, prognosis, risk, prediction
Hypothesis or research q.:	The intent is to identify key trends with respect to the types of machine learning methods being used, the types of training data being integrated, the kinds of endpoint predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or patient outcomes.
Summary:	In comparing and evaluating the existing studies a number of general trends were noted and a number of common problems detected. Some of the more obvious trends include a rapidly growing use of machine learning methods in cancer prediction and prognosis, a growing reliance on protein markers and microarray data, a trend towards using mixed (proteomic + clinical) data, a strong bias towards applications in prostate and breast cancer, and an unexpected dependency on older technologies such as artificial neural networks ANNs). Among the more commonly noted problems was an imbalance of predictive events with parameters (too few events, too many parameters), overtraining, and a lack of external validation or testing. Nevertheless, among the better designed and better validated studies it was clear that machine learning methods, relative to simple statistical methods, could substantially (15-25%) improve the accuracy of cancer susceptibility and cancer outcome prediction.
Significance (Novelty; importance, mechanistic insight)	In the past, our dependency on macro-scale information (tumor, patient, population, and environmental data) generally kept the numbers of variables small enough so that standard statistical methods or even a physician's own intuition could be used to predict cancer risks and outcomes. However, with today's high-throughput diagnostic and imaging technologies we now find ourselves overwhelmed with dozens or even hundreds of molecular, cellular and clinical parameters.

	In these situations, human intuition and standard statistics don't generally work. Instead we must increasingly rely on nontraditional, intensively computational approaches such as machine learning.
Critiques (data quality):	Summary of benefits, assumptions and limitations of different machine learning algorithms has been given in a beautiful manner. The details and various ways in which ML can be used in cancer prediction were included in the paper.
Future directions:	I observed that a brief introduction to machine learning was given in this paper. It helped in understanding the working of machine learning in cancer prediction and how Machine learning can be used to improve the current Cancer prediction models.

Research Paper-2

Complete citation:	**Breast Cancer Prediction system** Madhu Kumaria, Vijendra Singhb b a Department of Computer Science and Engineering, The NorthCap University, Sector 23A,Gurugram, Haryana, 122017, India (2018)
Key words:	WCBD, classification, knowledge mining, prediction system.
Hypothesis or research q.:	The intention of this study is to design a prediction system that can predict the incidence of the breast cancer at early stage by analyzing smallest set of attributes that has been selected from the clinical dataset. Wisconsin breast cancer dataset (WBCD) have been used to conduct the proposed experiment. The potential of the proposed method is obtained using classification accuracy which was obtained by comparing actual to predicted values. The outcome confirms that the maximum classification accuracy (99.28%) is achieved for this study
Summary:	A decision support system for predicting breast cancer helps and assist physician in making optimum, accurate and timely decision, and reduce the overall cost of treatment. Different classifiers have been used to conduct experiments on the standard WBCD. It is been observed KNN classifier yields the highest classification accuracies when used with most predictive variables. The proposed system greatly reduces the cost of treatment and improves the quality of life by predicting breast cancer at early stage of development.
Significance (Novelty; importance, mechanistic insight)	Many studies have been carried out, that uses number of different classifiers and approach for the prediction of breast cancer. Performance of their proposed method in this study shows best result as compared to the approach used by other authors on the same dataset
Critiques (data quality):	Good quality of data has been used in this research paper and helped in understanding the best classifier to predict breast cancer.
Future directions:	The future work will focus on exploring more of the dataset values and yielding more interesting outcomes. This study can help in making more effective and reliable disease prediction and diagnostic system which will

	contribute towards developing better healthcare system by reducing overall cost, time and mortality rate.

Research Paper-3

Complete citation:	**Lung cancer prediction using machine learning and advanced imaging techniques** Timor Kadir, Fergus Gleeson Optellum Ltd, Oxford, UK; Department of Radiology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK (2018)
Key words:	Pulmonary nodules; lung neoplasms; lung; machine learning; decision making
Hypothesis or research q.:	In this article, they have provided an overview of the main lung cancer prediction approaches proposed to date and highlight some of their relative strengths and weaknesses. They discuss some of the challenges in the development and validation of such techniques and outline the path to clinical adoption.
Summary:	Machine learning based lung cancer prediction models have been proposed to assist clinicians in managing incidental or screen detected indeterminate pulmonary nodules. Such systems may be able to reduce variability in nodule classification, improve decision making and ultimately reduce the number of benign nodules that are needlessly followed or worked-up
Significance (Novelty; importance, mechanistic insight)	They have provided an overview of the main approaches used for nodule classification and lung cancer prediction from CT imaging data. In their experience, given sufficient training data, the current state-of-the-art is achieved using CNNs trained with Deep Learning achieving a classification performance in the region of low 90s AUC points. When evaluating system performance, it is important to be aware of the limitations or otherwise of the training and validation data sets used, i.e., were the patients' smokers or nonsmokers, or were patients with a current or prior history of malignancy included.
Critiques (data quality):	Despite their classifier achieving the highest AUC and winning the competition, the performance was significantly below what they had seen on other independent datasets

Future directions:	They only used SVM classifier to classify and train the model. One critic could be a lack of independent training and validation data and avoid overfitting in future models. Try to achieve more accuracy for lung cancer prediction.

Research Paper-4

Complete citation:	**Lung Cancer Incidence Prediction Using Machine Learning Algorithms** Kubra Tuncal, Boran Sekeroglu, and Cagri Ozkan Near East University, Information Systems Engineering, Nicosia, TRNC, Mersin 10, Turkey (2020)
Key words:	lung cancer, support vector regression, backpropagation, long-short term memory
Hypothesis or research q.:	They predicted the cancer incidence rates of ten European countries (These countries are Germany, Denmark, Estonia, Finland, Iceland, Norway, Slovakia, Slovenia, Sweden and Geneva region of Switzerland), starting from 1970 to 2012 using machine learning models, Support Vector Regression, Long-Short Term Memory Network and Backpropagation Neural Network.
Summary:	In this paper, Support Vector Regression, Backpropagation Learning Algorithm and Long-Short Term Memory Network is used to perform lung cancer incidence prediction for ten European countries those records have been started from 1970. Results show that the prediction of incidence rates is possible with high scores with all algorithms; however, Support Vector Regression performed superior results than other considered algorithms.
Significance (Novelty; importance, mechanistic insight)	Lung cancer has highest mortality rates and this makes it more important to analyse the available data either it is insufficient. Lung cancer incidence rates of male and female data for ten European countries were analysed and prediction was performed using Support Vector Regression, Backpropagation and Long-Short Term Memory Network. Prediction results are analysed by using most efficient evaluation criteria in the literature; MSE, R2 and EV scores. Successful results were obtained for all algorithms; however, Support Vector Regression performed outstanding prediction results with the minimum error and the maximum prediction results. By considering other two algorithms, it was followed by Backpropagation and LSTM respectively.

Critiques (data quality):	They perform the predictions using three different algorithms (SVM, BA, LSTM) and provided the results in tabular format which was easily understandable.
Future directions:	Future work will include the implementation of more machine learning algorithms for the prediction of more cancer types for all European countries and dividing these predictions into age groups will be considered to analyse the incidence rates for age groups. Also, mortality rates will be included to predict both incidence and the mortality rates of patients

Research Paper-5

Complete citation:	**Lung Cancer Risk Prediction with Machine Learning Models** Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece (2022)
Key words:	healthcare; lung cancer; prediction; machine learning; data analysis
Hypothesis or research q.:	Here, a methodology for designing effective ML classification models is presented to predict lung cancer occurrence with the aid of the most common habits and symptoms/signs as input features to the models. Their contribution is a comparative assessment of numerous classifiers to develop the intended model with the highest sensitivity and discrimination ability in identifying those at high risk. For the evaluation of the models, they considered the performance metrics precision, recall, F-Measure, accuracy and AUC. Moreover, AUC ROC curves are also captured and presented. Finally, from various aspects, the performance analysis revealed that Rotation Forest is the most efficient model, and therefore constitutes the main proposition of this research article
Summary:	In this work, they used machine learning (ML) methods to build efficient models for identifying high-risk individuals for incurring lung cancer and, thus, making earlier interventions to avoid long-term complications. The suggestion of this article is the Rotation Forest that achieves high performance and is evaluated by well-known metrics, such as precision, recall, F-Measure, accuracy and area under the curve (AUC). More specifically, the evaluation of the experiments showed that the proposed model prevailed with an AUC of 99.3%, F-Measure, precision, recall and accuracy of 97.1%
Significance (Novelty; importance, mechanistic insight)	In the research article, for the topic under consideration, various ML models were employed in order to identify which one performs better than the rest by evaluating their prediction performance. Plenty of machine learning models, such as NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF and AdaBoostM1 are evaluated in terms of accuracy, precision, recall, F-Measure and AUC in order to determine the model with the best predictive performance.

Critiques (data quality):	They have worked on dataset publicly available and derived reliable and accurate research result.
Future directions:	To extend the current study along two axes. First, the more Supervised machine learning algorithm to use and comparing the results in terms of accuracy. Second, the evaluation of classification models in the same dataset will be made, an alternative data-splitting method for the models' validation, which applies resampling with replacement in the original data.

Research Paper-6

Complete citation:	**Lung Cancer Prediction from Text Datasets Using Machine Learning** BioMed Research International Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023
Key words:	Lung Cancer, prediction, Machine learning, KNN, data analysis.
Hypothesis or research q.:	In this paper, they optimise the process of detection in the lung cancer dataset using a machine learning model based on SVMs. Using an SVM classifier, lung cancer patients are classified based on their symptoms at the same time as the Python programming language is utilised to further the model implementation. There are various diagnostic methods for different tumours. But there are only a few specific ways to calculate what populations are in them.
Summary:	In this work, SVM is used to predict the development of lung cancer. The fundamental objective of this system is to provide consumers with an early warning, allowing them to save both money and time. The performance evaluation of the proposed method produced positive results, demonstrating that SVM can be used effectively by oncologists to aid in the identification of lung cancer. If the prediction is right, it is possible that the doctor will be able to prepare a better prescription and present the patient with an earlier diagnosis.
Significance (Novelty; importance, mechanistic insight)	This paper will introduce the method of not only diagnosing cancerous tumours but also doing the work required to calculate their size, shape, and location. Thus, not only can tumours be detected but also their type can be easily identified by counting and winning and can calculate the proper guidelines for dealing with them.
Critiques (data quality):	The procedures were mentioned in a well-defined manner, datasets ad results were also explained well.
Future directions:	There were some missing values in the dataset, which has an impact on the performance of the algorithm; therefore, caution should be exercised when analysing the data

OBJECTIVE:

- ✓ **Using larger Dataset:** Earlier all the authors tried with dataset ranging from 300 to 600 and they couldn't achieve best accuracy. Because, the more data you have for training, the more accuracy you get. That's why, I will be using the dataset consist of 1000 data. I have got the data from data.world website.

- ✓ **Resampling/Rearranging dataset:** Working on the same sequence of dataset might lead to less accuracy, so by changing the sampling ratio (random state) we can get better accuracy and model will be trained well.

- ✓ **Handling missing values:** Drawback in the previous proposed models were missing data in the dataset during training, which had affected the model to gain proper understanding of the data. So, I will be handling the missing data in my dataset to let me model be trained well on the datasets.

- ✓ **Change Proportion of Train and Test:** Splitting the dataset into different proportions, to get the best proportion for models' accuracy. Most used proportions are 75-25 / 80-20 / 90-10.

- ✓ **External validation on model:** Sometimes the model, that is being built, works best with known data but doesn't perform well with external data. So, to make model to predict accurate result for the data apart from dataset is significant. That's why, I will be feeding the model with unknown dataset to train the model on external data.

METHODOLOGY:

- ✓ **Import Libraries:** This step includes adding all the dependencies for the model to be trained. Such as various python libraries, modules and packages. Some of the vital libraries of python i.e. Sci-kit and Pandas will used in this Machine Learning project.

- ✓ **Data Collection:** All the data will be collected in this section and stored. My dataset is consisting of 1000 rows and 25 columns. Where each column contributes in some or the other way. So, I will be considering entire column for training my model.

- ✓ **Data Cleaning:** This is the step, where I will be handling the missing values, duplicated values or null values and try to overcome them for better model training and accuracy.

- ✓ **Data Analysis:** Data analysis is the essential step in every Machine learning project. Because, analysis help us to understand the data very well and get better knowledge about the data that we are working with.

- ✓ **Data Preprocessing:** The dataset we get is always an unfiltered data, so we need to filter them according to our requirements. If you train the data on unfiltered dataset, then your model will not be able to predict accurately for unknown data. Therefore, I will be performing tokenization, punctuation removal and stopwords removal, which will make my dataset more clear for training.

- ✓ **Data Visualization:** This step will help us to visualize our dataset, which will help us know more about the dataset.

- ✓ **Splitting data into training and testing set:** In this step, our dataset will be split into two parts i.e. training and testing and the proportion will be chosen based on the models performance.

- ✓ **Feature Extraction:** In this step, our dataset will be converted into the machines understandable format i.e. in a numerical format.

- ✓ **Train the model on training data:** The model will be trained in this step using various supervised and unsupervised learning models such as Support Vector Machine, Logistic Regression, Decision Tree and K-nearest neighbour.

- ✓ **Prediction on test data:** Now, I will be checking the model on test dataset for accuracy. Because, in some cases the model works well with training data but fails to predict for testing data and this situation is known as overfitting. And, for better model accuracy, we need to overcome the overfitting problem.

- ✓ **Check accuracy:** In this section, I will be checking the accuracy of the built model and various parameters such as Precision score, Accuracy Score and Confusion matrix will help me to do so.

- ✓ **Build predictive model:** Once the model achieves the best accuracy, we can build predictive model to predict on new datasets.