Prediction of the Incidence Rate of Lung Cancer Patients using Machine Learning

Anel Abdugulova[1], Aizhan Altaibek[1], Dina Kulzhanova[2] and Aigerim Altayeva[3]

[1] *International Information Technology University, Manas St. 34/1, Almaty, 050000, Kazakhstan*
[2] *National Pedagogical University named after Abai, Dostyk Ave 13, Almaty 050010, Kazakhstan*
[3] *Al-Farabi Kazakh National University, al-Farabi Avenue 71, Almaty, 050040, Kazakhstan*

Abstract

In this study considered one of the most common and deadly types of cancer, lung cancer presents substantial hurdles for early identification, prognosis, and therapy. In order to choose the best treatment options and enhance patient outcomes, accurate prediction of the lung cancer patient survival rate is essential. In this sense, machine learning approaches have shown to be quite promising. This research article gives a thorough investigation into the use of machine learning algorithms to forecast lung cancer patients' incidence rates. Various clinical and demographic factors, along with advanced imaging techniques, are employed as input features to develop predictive models. Various algorithms, including logistic regression, support vector machines (SVM), random forests, and artificial neural networks (ANN), are evaluated based on their performance metrics and predictive capabilities.

Keywords

Machine learning, cancer, lung cancer prediction, data preprocessing, neural network

1. Introduction

In recent years, the incidence of cancer and its consequences are increasing rapidly all over the world. The causes are many and complicated, but they take into account population aging and growth as well as shifts in the prevalence and distribution of the major cancer risk factors, many of which are linked to socioeconomic development [1, 2]. With the aging of the global population and the increasing rise of the cancer population, several nations have seen dramatic drops in the mortality rates of coronary heart disease and stroke in comparison to cancer.

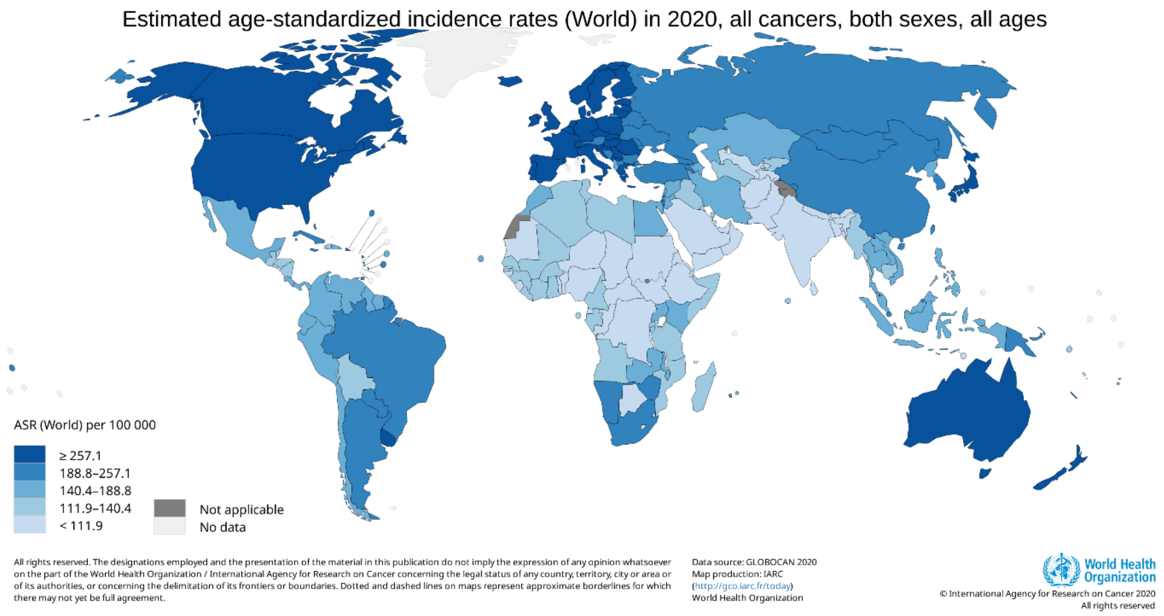
A worldwide estimated age-standardized incidence rates by 2020 are given in Figure 1. It covers all cancer types and includes both males and females in the age range of 0-74. The following data is available online at the Global Cancer Observatory [3].

Focusing on Asian countries, it becomes evident that Japan and the Republic of Korea exhibit the highest incidence rates, standing at 239.4 and 212.4, respectively. Meanwhile, Kazakhstan occupies the 11th position, reporting an age-standardized rate (ASR) of 150.3. Figure 2 shows that the incidence number in Kazakhstan is significant considering the size of the population.

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

anelabdugulova01@gmail.com (A. Abdugulova); aizhan.altaibek@yandex.ru (A. Altaibek); aikosha1703@gmail.com (A. Altayeva)

0009-0008-4516-8516 (A. Abdugulova); 0000-0001-8431-7950 (A. Altaibek); 0000-0002-9802-9076 (A. Altayeva)

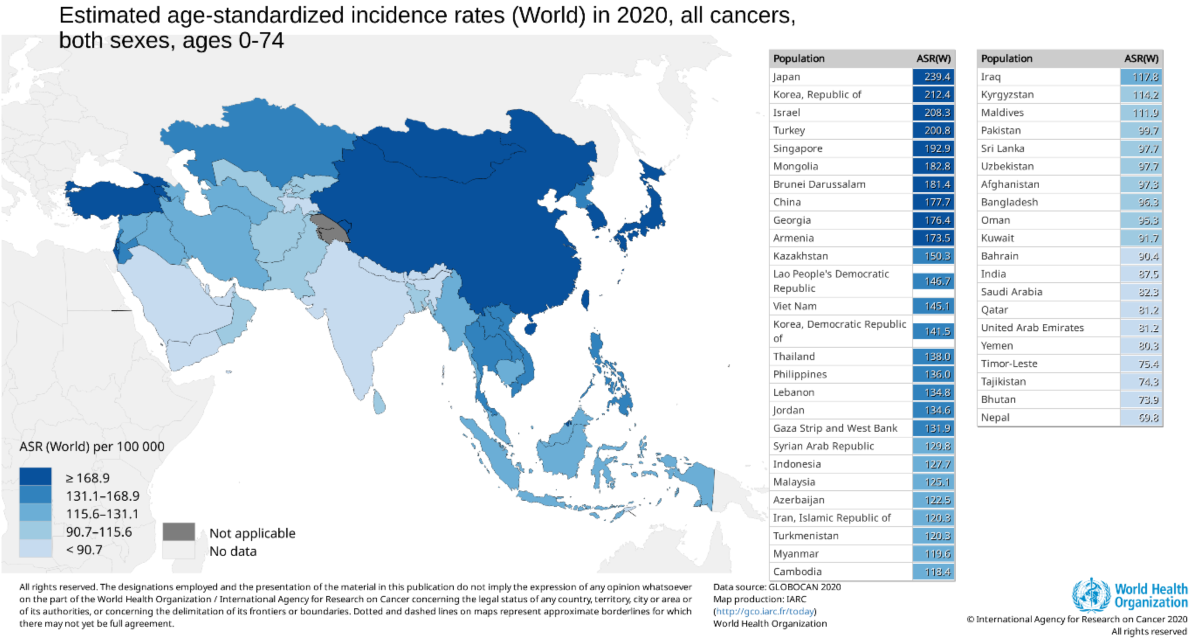

Estimated age-standardized incidence rates (World) in 2020, all cancers, both sexes, all ages
ASR (World) per 100 000
≥ 257.1
188.8–257.1
140.4–188.8
111.9–140.4
< 111.9
Not applicable
No data
World Health Organization

All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

Data source: GLOBOCAN 2020
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

© International Agency for Research on Cancer 2020
All rights reserved

Figure 1: Worldwide estimated age-standardized incidence rates in 2020


Estimated age-standardized incidence rates (World) in 2020, all cancers, both sexes, ages 0-74
ASR (World) per 100 000
≥ 168.9
131.1–168.9
115.6–131.1
90.7–115.6
< 90.7
Not applicable
No data
World Health Organization

Population	ASR(W)
Japan	239.4
Korea, Republic of	212.4
Israel	208.3
Turkey	200.8
Singapore	192.9
Mongolia	182.8
Brunei Darussalam	181.4
China	177.7
Georgia	176.4
Armenia	173.5
Kazakhstan	150.3
Lao People's Democratic Republic	146.7
Viet Nam	145.1
Korea, Democratic Republic of	141.5
Thailand	138.0
Philippines	136.0
Lebanon	134.8
Jordan	134.6
Gaza Strip and West Bank	131.9
Syrian Arab Republic	129.8
Indonesia	127.7
Malaysia	125.1
Azerbaijan	122.5
Iran, Islamic Republic of	120.3
Turkmenistan	120.3
Myanmar	119.6
Cambodia	118.4

Population	ASR(W)
Iraq	117.8
Kyrgyzstan	114.2
Maldives	111.9
Pakistan	99.7
Sri Lanka	97.7
Uzbekistan	97.7
Afghanistan	97.3
Bangladesh	96.3
Oman	95.3
Kuwait	91.7
Bahrain	90.4
India	87.5
Saudi Arabia	82.3
Qatar	81.2
United Arab Emirates	81.2
Yemen	80.3
Timor-Leste	75.4
Tajikistan	74.3
Bhutan	73.9
Nepal	69.8

All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

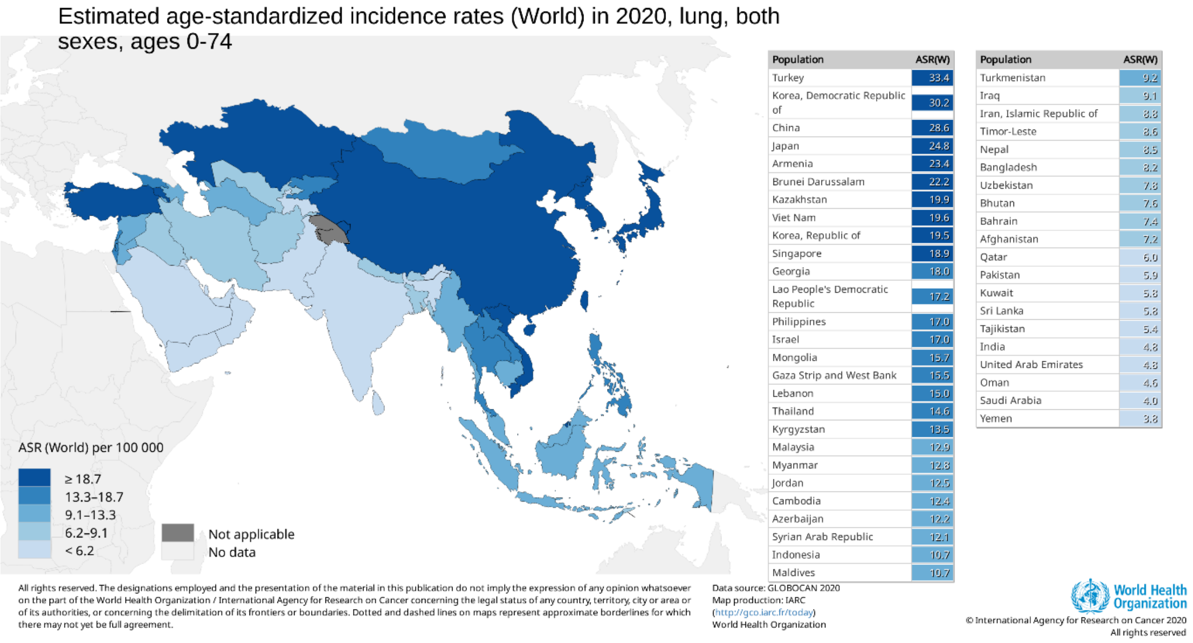
Data source: GLOBOCAN 2020
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

© International Agency for Research on Cancer 2020
All rights reserved

Figure 2: Estimated age-standardized incidence rates of Asian countries in 2020 [4]

This study solely looks at research on lung cancer, one of the most widespread and common cancers in the world, out of the more than 100 different types of cancer. Lung cancer has remained one of the most important medical and socioeconomic problems in recent [5]. It is mainly caused by air pollution, cigarette smoking, and cardiopulmonary syndrome, detected by a statistically significant association [6, 7].

Every year, this pathology takes the lives of up to 1.8 million people. More than 2 million men and women are diagnosed with lung cancer each year, with men making up two-thirds of those cases (1,368,524) and women making up one-third (725,352), according to the International Agency for Research on Cancer (IARC). Based on the statistics of lung cancer incidence by 2020 (Figure 3) Kazakhstan is among the first ten countries topping the list.


Estimated age-standardized incidence rates (World) in 2020, lung, both sexes, ages 0-74
ASR (World) per 100 000
≥ 18.7
13.3–18.7
9.1–13.3
6.2–9.1
< 6.2
Not applicable
No data
World Health Organization

Population	ASR(W)
Turkey	33.4
Korea, Democratic Republic of	30.2
China	28.6
Japan	24.8
Armenia	23.4
Brunei Darussalam	22.2
Kazakhstan	19.9
Viet Nam	19.6
Korea, Republic of	19.5
Singapore	18.9
Georgia	18.0
Lao People's Democratic Republic	17.2
Philippines	17.0
Israel	17.0
Mongolia	15.7
Gaza Strip and West Bank	15.5
Lebanon	15.0
Thailand	14.6
Kyrgyzstan	13.5
Malaysia	12.9
Myanmar	12.8
Jordan	12.5
Cambodia	12.4
Azerbaijan	12.2
Syrian Arab Republic	12.1
Indonesia	10.7
Maldives	10.7

Population	ASR(W)
Turkmenistan	9.2
Iraq	9.1
Iran, Islamic Republic of	8.8
Timor-Leste	8.6
Nepal	8.5
Bangladesh	8.2
Uzbekistan	7.8
Bhutan	7.6
Bahrain	7.4
Afghanistan	7.2
Qatar	6.0
Pakistan	5.9
Kuwait	5.8
Sri Lanka	5.8
Tajikistan	5.4
India	4.8
United Arab Emirates	4.8
Oman	4.6
Saudi Arabia	4.0
Yemen	3.8

All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

Data source: GLOBOCAN 2020
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

© International Agency for Research on Cancer 2020
All rights reserved

Figure 3: Estimated age-standardized lung cancer incidence rates of Asian countries in 2020 [8]

Scientists and practitioners attend every event devoted to this issue as a result, to share knowledge and find creative ideas to help the lung cancer condition [9]. The presence of such problems requires the solution to a various task such as designing cancer risk-prediction models. Models try to determine cases with a higher risk of cancer development than the general incidents and study the progression of the disease to improve survival rates, and build methods that trace the effectiveness of treatment to improve treatment options [10-12].

In recent years Machine learning (ML) has grown rapidly due to data collection improvements, capacity processing, and algorithmic innovation [13]. ML techniques are essential for resolving a wide range of complicated issues in a variety of industries beginning with healthcare and banking to image recognition and even natural language processing. It's crucial to comprehend the advantages and disadvantages of various machine learning approaches in order to choose the one that is best suited for a given task.

To assist researchers, data scientists, and practitioners in selecting the best tool for their particular applications, this study will examine and contrast some well-known machine learning approaches in this article [14].

2. Background

Lung cancer initiation takes place within the pulmonary tissue and may potentially metastasize to other anatomical regions. The pathogenesis is attributed to the proliferation of malignant cells within the lung tissue. Lung cancer is notably categorized into two primary forms, namely non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). These subtypes exhibit variations in their etiological pathways and therapeutic modalities. The leading causative factor in the development of lung cancer is cigarette smoking, predominantly associated with tobacco consumption. Nevertheless, it is imperative to acknowledge that lung cancer can also manifest in non-smokers [15, 16]. This ailment is recognized as a substantial global public health concern, particularly evident in regions such as Japan and South Korea, where the incidence rates are notably elevated.

A pivotal investigation titled "Advancements in Lung Cancer Screening and Early Detection" conducted two randomized controlled trials, which established that the utilization of low-dose computed tomography (LDCT) for screening results in a statistically significant reduction in mortality among lung cancer patients. Consequently, LDCT has become the prevailing standard for lung cancer screening.

Nevertheless, numerous issues persist in the domain of lung cancer screening, encompassing topics such as screening criteria, a relatively high incidence of false-positive outcomes, and the concern of radiation exposure.

In recent years, the field of machine learning (ML) has witnessed substantial progress, with a multitude of research endeavors aiming to prognosticate diverse forms of cancer. The present article introduces a diverse array of ML models, including but not limited to logistic regression, decision trees, K Nearest Neighbors, Random Forests, Support Vector Machines, Neural Networks, and CyPath. It comprehensively outlines the application domains and the performance characteristics of these models in the context of cancer prediction and diagnosis.

3. Data collection and preprocessing

In order to get higher accuracy and the best results a comprehensive dataset comprising clinical, demographic, and imaging data of cancer patients is needed. The data undergoes rigorous preprocessing steps, including data cleaning, feature selection, and normalization, to ensure the quality and suitability for model development. The result of preprocessing will be a training set suitable for data training [17].

1. Data cleaning

The first step is to get the most meaningful information since the occurrence of noisy data as null values and outlying data can significantly impact prediction results [18]. Cleaning of character data includes steps such as Tokenization - identifying different words. The second step is morphological transformations, which aims to convert words into the base form. There are two ways to make a transformation: Stemming and Lemmatization. The last step is removing useless text [19].

Data cleaning also includes de-duplication and correction of misspelled string. The study “Data Cleaning: Current Approaches and Issues” gives examples of existing data cleansing techniques [20]. As a result of numerous numbers of research, it is obvious that data set noise significantly impacted classification accuracy and produced bad predictions [18].

2. Feature selection

According to the review of foreign literature, much attention is paid to the risk factors of lung cancer, which is also important for assessing the prediction of survival. Presented below is a compilation of valuable attributes and factors for the detection of lung cancer:

1. Smoking History will be one of the most important features since it is a leading cause of lung cancer, and a long history of tobacco use is a significant risk factor.
2. Lung cancer risk increases with age, and older individuals are more likely to develop the disease.
3. A family history of lung cancer can elevate an individual's risk.
4. Occupational exposure to carcinogens like asbestos, radon, or certain chemicals can increase the risk of lung cancer.
5. Exposure to ionizing radiation, such as during medical treatments, can contribute to lung cancer risk.
6. Men historically have had a higher risk of lung cancer, although this gap has been closing.
7. Long-term exposure to air pollutants, especially in urban areas, may increase lung cancer risk.
8. A diet high in fruits and vegetables, as well as low in red meat may reduce the risk of lung cancer.
9. Regular physical activity and a healthy lifestyle, which can lower the risk of cancer.
10. Conditions like chronic obstructive pulmonary disease (COPD) and emphysema.
11. The presence of suspicious lung nodules on imaging studies as CT scans, may indicate early-stage lung cancer.

12. Cough and Respiratory Symptoms: Persistent cough, changes in cough patterns, and other respiratory symptoms can be warning signs.

13. Examination of sputum (mucus from the lungs) for abnormal cells.

14. Certain genetic mutations like those in the EGFR gene are associated with a higher risk of lung cancer and can be targeted for treatment.

15. Specific proteins or substances in the blood as progastrin-releasing peptide (ProGRP) or carcinoembryonic antigen (CEA) can serve as biomarkers for lung cancer.

16. High-resolution computed tomography (CT) scans and positron emission tomography (PET) scans.

17. Bronchoscopy allows direct visualization of the airways and can aid in the diagnosis of lung cancer.

18. Biopsy by obtaining tissue samples can confirm the presence of cancer and determine its type and stage.

19. Liquid Biopsies which is an emerging technology that allows the detection of cancer-related genetic mutations and biomarkers in blood samples.

20. Advanced algorithms and machine learning models are being developed to analyze medical images and identify lung cancer with high accuracy.

As stated in the 16th point CT is highly effective in providing precise diagnoses of lung diseases when compared to simple chest radiography. CT offers superior image resolution at a lower cost compared to magnetic resonance imaging (MRI).

However, one significant drawback of traditional or "normal-dose" CT scans is that they expose patients to high levels of radiation, which raises ethical concerns due to potential health risks. This has led to limitations on the use of normal-dose CT scans.

To address this concern, the concept of low-dose CT (LDCT) was introduced. LDCT scans have gained attention because they significantly reduce radiation exposure, making them safer for patients. LDCT scans are approximately three times less harmful than normal-dose CT scans, while still allowing the detection of microscopic diseases that might not be visible in standard X-ray images. Consequently, LDCT has become a popular choice as the initial screening test for identifying lung anomalies due to its high sensitivity.

LDCT may have a higher false-positive rate compared to normal-dose CT scans when diagnosing lung diseases. This can be attributed to factors such as increased image noise, artifacts, or difficulties in image reconstruction caused by the lower radiation dose used in LDCT imaging.

3. Normalization

The third pre-processing approach is data normalization, which consists of removing redundant or unstructured data. Normalization aims to validate logical data storage and to make an equal contribution of each feature. The success of machine learning algorithms depends upon the quality of the data to obtain a generalized predictive model of the classification problem [21].

Many authors have validated the impact of data normalization for improving classification performance in various fields, such as medical data classification [22].

Previous studies on normalization compared small networks and complex ones on the same network architecture. Comparison shows that it is better to use small than more complex networks [23].

4. Methods and research or Machine learning techniques

Linear Regression

Linear regression is a fundamental technique for modeling the relationship between a dependent variable and one or more independent variables. It is a widely employed statistical technique in the realm of medical research for predicting the likelihood of cancer occurrence in individuals. Tolles J and Meurer WJ. Their study points out that Logistic regression is also applicable to multinomial regression problems [24]. This method is particularly valuable for binary classification tasks, such as distinguishing between cancer and non-cancer cases. Logistic regression models the probability of an event (in this case, the presence of cancer) based on one or more predictor variables, which may encompass various clinical, demographic, or diagnostic factors.

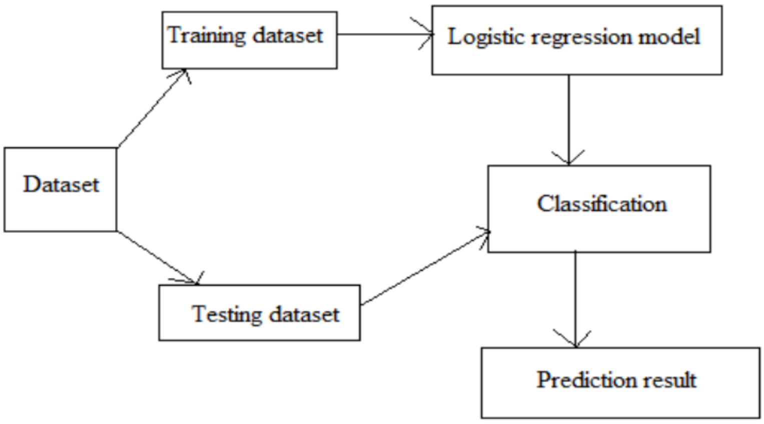


Figure 4: Steps involved in constructing logistic regression model [24]

In Figure 4 above given a sequence of using logistic regression with a preprocessed dataset. First step is dividing dataset into two subsets: training and testing. Then logistic regression is applied to the training part of the dataset and both subsets are classified. In the last step predicted values are compared to testing dataset results and model accuracy is calculated.

In the results of the study of Mr. Raghavendra Patil G. E., Ms. Sinchana C. G. and Ms. Tejashwini P., “Lung cancer prediction system using logistic regression approach” training accuracy was 96% and testing accuracy was 85%. The survivability rate of lung cancer can be predicted with the help of modern machine learning techniques like logistic regression where the proposed system can predict the lung cancer in the early stages which helps the survivability rate of the patients.

In the study of Lavanya C., Pooja S., and Abhay H. Kashyap the validation sets gave the following results (weighted average): accuracy of 74%, precision of 75%, recall of 75%, and F1 score of 75%. For Logistic regression, the size of the parameter space is small, and thus an exhaustive search for all the combinations was performed using GridSearchCV() [25].

K-Nearest Neighbors (K-NN)

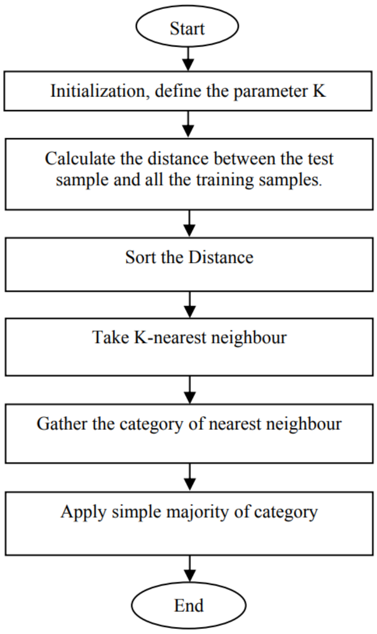

Start
Initialization, define the parameter K
Calculate the distance between the test sample and all the training samples.
Sort the Distance
Take K-nearest neighbour
Gather the category of nearest neighbour
Apply simple majority of category
End

Figure 5: Implementation steps of K-NN algorithm

K Nearest Neighbors Classification(K-NN) is one of the most common machine learning classifiers, which is easy to understand visually. KNN works well for clustering and recommendation systems but for large datasets can be computationally expensive.

From the perspective of pattern recognition, the K-NN algorithm is a non-parametric method used for classification and regression. In the feature space, the K-number of the closest training samples forms an Input, and a class membership forms an output. If K = 1, then the class is single nearest neighbor [26]. K-NN is a type of instance-based learning.

Figure 5 represents the steps of implementation of the K-NN algorithm.

In the study by Lavanya C., Pooja S., Abhay H. Kashyap the validation sets gave the (weighted average): accuracy of 79%, precision of 75%, recall of 82%, and F1 score of 77%. [25].

The KNN can rival the most reliable models since it makes exceptionally precise forecasts (Lu, Zhu, and Gu 2014). Consequently, you can involve the KNN calculation for applications that require high precision yet that don't need a comprehensible model.

Decision Trees

A decision support tool known as a decision tree employs a tree-like graph or model to represent decisions and all possible outcomes, such as utility, resource costs, and chance event outcomes. It is also considered as one of the approaches to display an algorithm. In a decision tree, every internal node can be considered a "test" on an attribute (e.g. a coin flip), each branch reflects the result of the test, and each leaf node represents a class label (the choice made after computing whether a coin will land on its head or tail) [27].

Decision trees are universal and interpretable models, often used for classification and regression tasks by separating data based on attributes. This makes them well suited for tasks where the importance of functions is crucial. But despite these advantages, decision trees are unfortunately subject to retraining.

There are a few steps for the construction of a Decision tree:

1. Check whether all the cases belong to the same class and if “Yes” then the tree is a leaf and that node is labeled by that class.

2. The increment of information is calculated for each attribute

3. Accept the best selection criteria and, accordingly, consider the separation attribute.

4. Count the information gain: The concept of entropy arrives in this part. Entropy can be stated as its measure of any disorder in the data. Entropy can also be called as a measurement of uncertainty in any random variable.

5. Pruning: For the tree creation process, pruning is an important technique to be performed.

A dataset can sometimes contain subsets that are not well-defined instances, so pruning can be used to classify such subsets.

6. Pruning is of two types:

1. Subsequent pruning, which is performed after the tree is created.

2. Operational pruning, performed in the process of creating a tree.

The novel Decision tree is one more administered learning calculation that goes under the characterization strategy. The novel Decision tree is a plan, where every inside center point meaning a test set branch tends to the consequences of nodes and test leaf center points contain the aftereffect of the class marks. It has three estimating boundaries that are data gain, gain proportion, and gini file (Kubík and Polák 1986)

According to the results of the study “Decision Tree Over KNN For Lung Cancer Detection to Increase Accuracy” the novel decision tree algorithm is better than the KNN algorithm. From the lung cancer dataset, it is verified that the accuracy of the decision tree was much greater than the k-nearest neighbor algorithm.

The Novel Decision Tree has an accuracy of 98.06% which is higher than the KNN which has 90.73% [28].

Random Forests

Random Forests address decision tree overfitting by aggregating multiple decision trees into an ensemble model [29]. This technique excels in reducing variance and increasing accuracy and is used in image classification, bioinformatics, and fraud detection.

An example of an ensemble learning technique is a random forest classifier, which is made up of many decision trees. It has a straightforward framework that is easy to comprehend and more effective than other methods. The capacity to adapt to problem space settings and independence from the data domain is the most demanding concern with different types of classifiers. Consequently, the Random Forest classifier performs better by classifying a data point and combining the prediction of multiple trees.

For training the Machine Learning model for the dataset in the study "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers" by Lavanya C., Pooja S., Abhay H. Kashyap, Abdur Rahaman, the following features were selected, FC, logFC, and P-value.

The best results were obtained with n scores=1200, maximum characteristics=sqrt, minimum sample sheet=4, n tasks=-1 and maximum depth=100. In this case, the training and validation sets gave the following results: accuracy of 87%, precision of 86%, recall of 84% and F1 score of 85%. helped us in classifying the biomarkers causing non-small cell lung cancer and small cell lung cancer [25].

Support Vector Machines (SVM)

SVMs are efficient for binary classification and can process nonlinear data by mapping it into a multidimensional space. They are known for their high generalization performance but may require significant computational resources for large datasets.

The classifier of machine support vectors works by finding the decision boundary in such a way that the data points during construction are divided into classes by this hyperplane in the object space, and the separation is as possible. Thus, it is called the maximum margin classifier [30].

In the study, GridSearchCV() is used to perform this hyperparameter tuning. The best results were given when C=1000 and gamma=1, with the RBF kernel. In this case, the validation sets gave the following results: accuracy of 81%, precision of 83%, recall of 82% and F1 score of 82% [25].

Usage of CyPath

In the development of the CyPath Lung cancer/non-cancer classifier, a comprehensive approach was taken to enhance the early detection of lung cancer using sputum samples [31]. The key findings and steps include:

Logistic Regression Modeling: The CyPath Lung assay uses logistic regression models to classify sputum samples as either cancer or non-cancer. These models are based on a set of predictor variables and aim to predict the binary response variable (cancer or non-cancer) [32].

Predictive Variables: Age emerged as a significant clinical parameter in the model, consistent with its established relevance to lung cancer. Other factors, such as smoking history and specific fluorescence signal densities (TCPP/log10SSC-A and FVS510-A/log10FSC-A), were identified as informative predictors for the classifier.

Running the CyPath Lung Assay Pipeline: The assay's pipeline involves several stages, including quality control checks, the determination of predictive variable values, and sample classification. The classifier uses age and flow-based values from viable singlet cells to assess the likelihood of cancer.

Performance of CyPath Lung: The CyPath Lung assay exhibited robust performance metrics. It demonstrated high sensitivity, specificity, and accuracy for samples analyzed on both the LSRII and Navios EX flow cytometers. The negative predictive value (NPV) exceeded 95%, indicating its reliability in correctly identifying non-cancer samples [33].

Performance in Smaller Nodules: Notably, the analysis performed exceptionally well in cases with smaller lung nodules, with a sensitivity of 92% and specificity of 87%. This suggests its effectiveness in identifying lung cancer in individuals with less advanced diseases [34]. Model Contribution: Each predictor retained in the model significantly contributed to its overall performance, reinforcing the importance of considering multiple factors in lung cancer prediction.

In summary, the CyPath Lung assay represents a promising advancement in the early detection of lung cancer. [35] It leverages logistic regression modeling and a combination of demographic and flow-based variables to achieve strong accuracy, especially in identifying cancer in individuals with smaller lung nodules. This approach holds the potential for improving lung cancer diagnosis, particularly in high-risk individuals undergoing screening. Age, while a significant factor, is not the sole determinant of the model's success, emphasizing the importance of a multifaceted approach to prediction [35].

Neural Networks:

Neural networks, particularly deep learning models, have achieved remarkable success in image and speech recognition, natural language processing, and autonomous systems. Their deep architectures can capture intricate patterns in data but often require extensive computational resources and substantial amounts of labeled data.

ANNs have unique properties including robust performance in dealing with noisy or incomplete input patterns, high fault tolerance, and the ability to generalize from the training data [36]. Deep neural networks, or DNNs, are widely used in several fields and produce output based on input variables. DNNs are inspired by the way the brain works. Based on a target and a set of features, DNNs can train to construct nonlinear function approximations. Hidden layers are those that are positioned between the input and output layers. The deep neural network (DNN) is capable of learning intricate nonlinear function relationships from high-dimensional raw data without the use of artificial rules, thanks to its numerous nonlinear hidden layers [37].

They proposed an ANN to predict the category of movie rate [38], predict the price range of mobile phones [39], predict the category of animals [40], diagnose the category of tumors [41], and diagnose Autism [42]. ANN model was able to predict the presence of lung cancer with 96.67% accuracy, after 1418105 learning cycles with less than 1% training error rate as seen in figure (3). In addition, the Model showed that the most attribute that affects the lung cancer presence is age [43].

The highest classification accuracy of 99.7% for lung cancer classification was reported by work in [43]. The Discrete AdaBoost Optimized Ensemble Learning Generalized Neural Network (DAELGNN) framework, which distinguishes between non-normal (cancerous) and normal lung features using a set of normalized biological data points, was developed in this study.

5. Results and discussion

Machine learning techniques have become instrumental in solving complex problems across different industries, including healthcare, finance, image recognition, and natural language processing.

The choice of a machine learning technique should be driven by the specific requirements of a given problem, the available data, and computational resources.

Each technique has its unique characteristics, strengths, and weaknesses, so it is important to choose the most suitable for the task. Therefore, it's essential to perform thorough model selection and evaluation to ensure the chosen technique aligns with the problem's objectives.

Logistic regression, for example, is effective in distinguishing between benign and malignant breast lesions. The CyPath Lung assay uses logistic regression to detect lung cancer early, especially in smaller nodules.

Decision trees are versatile but can overfit, while K Nearest Neighbors excels in high-precision applications. Random Forests improve upon decision trees' limitations. Support Vector Machines handle non-linear data but require substantial computational resources. Neural Networks, especially deep learning models, are successful in tasks like image recognition.

6. Conclusion

Cancer incidence and mortality rates are increasing globally due to factors like population aging, shifting cancer risk factors, and socioeconomic development. Some countries are witnessing more cancer-related deaths than those caused by coronary heart disease and stroke. A closer look at Asian countries reveals high cancer incidence rates, particularly in Japan and South Korea.

Among various cancer types, lung cancer remains a significant global health issue with strong associations with risk factors like air pollution and smoking.

Machine learning is a dynamic field, and ongoing research leads to the development of new algorithms and improvements in existing ones. Staying updated with the latest advancements and being flexible in adapting the most suitable technique for each task is essential for success in the world of machine learning. As the field of machine learning advances, it offers valuable tools for addressing complex cancer-related tasks, from risk prediction models to treatment effectiveness assessments.

The comparative analysis provided here serves as a starting point for selecting the right machine learning approach, but ongoing exploration and experimentation are key to achieving optimal results in diverse applications.

Based on the results of this comparison in this study, we can say that many studies have been conducted on lung cancer prediction and very impressive results have been obtained. However, it is worth noting that all models show good results and that the further development of this area (ML) will have a very good effect on mortality and earlier detection of lung cancer.

The study requires further research to explore the interaction between the latest cancer detection methods and computational models to improve prediction accuracy in detecting lung cancer in the early stages.

7. References

[1] Abdel R. Omran. (1971). The epidemiologic transition: A theory of the epidemiology of population change. Milbank Mem Fund Q. 49. pp. 509-538.

[2] Gersten O., Wilmoth J.R. (2002). The cancer transition in Japan since 1951. Demogr Res. 7. pp.271-306.

[3] International Agency for Research on Cancer 2023. URL: https://gco.iarc.fr/today/online-analysis.

[4] International Agency for Research on Cancer 2023. URL: https://gco.iarc.fr/today/online-analysismap?v=202.

[5] Kazakhstanskiy pfarmatsevticheskiy vestnik. URL: https://pharmnewskz.com/ru/article/rak-legkogo-peredovye-resheniya_18263.

[6] Vital, T., Panduranga. (2014). Data collection, statistical analysis and clustering studies of cancer dataset from viziayanagaram District, AP, India. ICT and critical infrastructure. In: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II. Springer, Cham.

[7] Douglas, P.K., Harris, S., Yuille, A., Cohen. (2011). Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief versus disbelief. Neuroimage. 56(2), pp.544–553.

[8] International Agency for Research on Cancer 2023. URL: https://gco.iarc.fr/today/online-analysis-map?v=2020.

[9] Kazakhstanskiy pfarmatsevticheskiy vestnik. URL: https://pharmnewskz.com/ru/article/rak-legkogo-peredovye-resheniya_18263.

[10] Kourou K., Exarchos T.P., Exarchos K.P. (2015). Machine learning applications in cancer prognosis and prediction." Comput Struct Biotechnol J. 13. pp.8–17. doi: 10.1016/j.csbj.2014.11.005.

[11] Iqbal M.J., Javed Z., Sadia H. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int. 21(1). pp.1–11. doi: 10.1186/s12935-021-01981-1.

[12] Loud J.T., Murphy J. (2017). Cancer screening and early detection in the 21st century. Semin Oncol Nurs. 33. pp. 121–128. Doi 10.1016/j.soncn.2017.02.002.

[13] George Tzanis, Ioannis Katakis and Ioannis Partalas. Modern Applications of Machine Learning. Department of Informatics, Aristotle University of Thessaloniki, GR-54124.

[14] Caichen Li, Huiting Wang and Yu Jiang. (2022). Advances in lung cancer screening and early detection. Cancer Biol Med. 19(5). pp. 591–608.

[15] Bade B.C., Cruz C.S.D. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. Clin Chest Med. 41(1). Pp. 1–24. doi: 10.1016/j.ccm.2019.10.001.

[16] Barta J.A., Powell C.A. and Wisnivesky J.P. (2019). Global epidemiology of lung cancer. Ann Global Health. 85:1. doi: 10.5334/aogh.2419.

[17] S.B. Kotsiantis, D. Kanellopoulos and P.E. Pintelas. (2006). Data Preprocessing for Supervised Learning. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE. 1(1). ISSN 1306-442.

[18] Shivani Guptaa, Atul Guptab. (2019). Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. Proceedings of the Fifth Information Systems International Conference 2019, Manipal University Jaipur, India.

[19] Ahmad Sadek. (2022). Primary stage Lung Cancer Prediction with Natural Language Processing-based Machine Learning" Stockholm, Sweden, 2022. URL: https://www.diva-portal.org/smash/get/diva2:1676617/FULLTEXT01.pdf.

[20] Vaishali C. W. and Ratnadeep R.D. Data Cleaning: Current Approaches and Issues.

[21] Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97 (Part B), December 2020, 105524. URL: https://www.sciencedirect.com/science/article/abs/pii/S1568494619302947.

[22] H. Hannah Inbarani, Ahmad Taher Azar and G. Jothi. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. doi: 10.1016/j.cmpb.2013.10.007.

[23] J. Sola and J. Sevilla. Importance of input data normalization for the application of neural networks to complex industrial problems. Department of Electrical and Electronic Engineering, Universidad Pública de Navarra. 31006 Pamplona, Spain. doi::10.1109/23.589532.

[24] Tolles J., Meurer W.J. (2015). Logistic regression. JAMA. pp. 316:533.

[25] Lavanya C., Pooja S. and Abhay H. Kashyap. Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. Department of Biotechnology, RV College of Engineering, Bengaluru, Karnataka, India. doi:10.1177/11769351231167992. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm (2010). URL: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/.

[26] Ms. Leena Patil, Ms. Aparna Sirsat and Ms. Diksha Kamble. (2017). Lung Cancer Detection using Decision Tree Algorithm. International Research Journal of Engineering and Technology (IRJET). 4(02).

[27] N. Charan and S. Parthiban. (2019). Decision Tree Over KNN For Lung Cancer Detection to Increase Accuracy. Journal of Survey in Fisheries Sciences. 10(1S). pp.2934-2943.

[28] Vrushali Y Kulkarni and Dr Pradeep K Sinh. Random Forest Classifiers: A Survey and Future Research Directions. International Journal of Advanced Computing. 36(1). ISSN:2051-0845.

[29] Cortes C. and Vapnik V. (1995). Support-vector networks. Mach Learn. 20. pp.273-297.

[30] Shi H.Y., Hwang S.L., Lee K.T., et al. (2013). In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. J Neurosurg. 118. Pp.746-52.

[31] Park Hyeoun-Ae. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. J Korean Acad Nurs. 43(2). No.2, pp.154-164. URL:https://jkan.or.kr/DOIx.php?id=10.4040/jkan.2013.43.2.154.

[32] Schmidhuber J. (2015). Deep learning in neural networks: an overview. Neural Netw. 61. Pp.85–117.

[33] Nasser, I. M. and Abu-Naser, S. S. (2019). Artificial Neural Network for Predicting Animals Category. International Journal of Academic and Applied Research (IJAAR), 3(2).

[34] Precision Pathology Laboratory. URL: https://www.precisionpath.us/services/cypath-lung-faq/2.

[35] Brameier, M. (2009). Data Mining and Knowledge Discovery, Neural Networks in. In: Meyers, R. (eds) Encyclopedia of Complexity and Systems Science. Springer, New York, NY. https://doi.org/10.1007/978-0-387-30440-3_116.

[36] Nasser, I. M., Al-Shawwa, M., & Abu-Naser, S. S. (2019). Artificial Neural Network for Diagnose Autism Spectrum Disorder. International Journal of Academic Information Systems Research (IJAISR), 3(2).

[37] Nasser, I. M., Al-Shawwa, M., & Abu-Naser, S. S. (2019). A Proposed Artificial Neural Network for Predicting Movies Rates Category. International Journal of Academic Engineering Research (IJAER), 3(2).

[38] Nasser, I. M., Al-Shawwa, M., & Abu-Naser, S. S. (2019). Developing Artificial Neural Network for Predicting Mobile Phone Price Range. International Journal of Academic Information Systems Research (IJAISR), 3(2).

[39] Nasser, I. M., & Abu-Naser, S. S. (2019). Artificial Neural Network for Predicting Animals Category. International Journal of Academic and Applied Research (IJAAR), 3(2).

[40] Nasser, I. M., Al-Shawwa, M., & Abu-Naser, S. S. (2019). Artificial Neural Network for Diagnose Autism Spectrum Disorder. International Journal of Academic Information Systems Research (IJAISR), 3(2).

[41] Ibrahim M. Nasser, Samy S. Abu-Naser. Lung Cancer Detection Using Artificial Neural Network". International Journal of Engineering and Information Systems (IJEAIS), 3(3).

[42] Shakeel PM, Tolba A, Al-Makhadmeh Z, Jaber MM. (2020). Automatic detection of lung cancer from biomedical data set using discrete adaboost optimized ensemble learning generalized neural networks. Neural Comput Appl. 32(3). Pp.777–790. doi: 10.1007/s00521-018-03972-2.