JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024
Research Article
https://doi.org/10.56979/702/2024

Multi-Model Machine Learning Analysis of Environmental Risk Factors for Lung Cancer

Naeem Abbas[1], Muhammad Azam[2], Muazzam Ali[1*], Mafia Malik[1], M U Hashmi[1,] and Abdul Manan[1]

[1]Deparment of Basic Sciences, Superior University, Lahore, 54000, Pakistan.
[2]Deparment of Computer Science, Superior University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Muazzam Ali. Email: muazzamali@superior.edu.pk

Received: April 11, 2024 **Accepted:** August 29, 2024 **Published:** September 01, 2024

Abstract: One sort of malignant growth that catches the lungs is cellular breakdown in the lungs(LC). It is one of the main sources of mortality in the modern world. The smoke created by the deficient ignition of biomass fuels contains different unsafe synthetic compounds or chemicals that can be incredibly risky to human health. Around 25% of examples of cellular breakdown in the lungs universally not connected to tobacco use. The genomic scene of cellular breakdown in the lungs likewise incorporates modifications to DNA repair pathways, hereditary genetic risk variables, and variation in gene expression. Air pollution, toxins, and tobacco smoke are a couple of representation of ecological factors that extraordinarily lift the danger of cellular breakdown in the lungs, even while hereditary factors remain a major influence on lung cancer susceptibility and progression. There's no believable exploration that could give data about Pakistan's ongoing indicative strategies. The prediction and early identification of cellular breakdown in the lungs save endless lives. Accordingly, strong machine learning algorithms calculations are expected to distinguish event of LC in its beginning phases. Perceiving the different characters of abnormal growth of cells in the lungs etiology, this study focuses on various ecological causes including air contamination, tobacco smoke, exposure to radiation and hereditary inclination. Models created utilizing ML algorithms like SVM, KNN, and NB etc. As by all the premier accuracy obtained by the classifier DT which is 99.67. In our review, we additionally endeavoured to reveal relationships between the different elements in the dataset utilizing traditional machine learning approaches. Clinical specialists in their facilities can involve this model as a choice help framework.

.

Keywords: Genetics; Smoking; Cellular breakdown in the Lungs or Lung Cancer; Air Contamination or Pollution; Machine Learning Models.

1. Introduction

The The lungs, situated in the chest, are wipe like organs that are divided into segments known as curves. Air goes through the trachea, bronchi, and bronchioles during relaxing. Alveoli are little air sacs situated toward the finish of the bronchioles[1,2]. Disease is a gathering of distorted cells that create and spread to different tissues. The human body comprises of trillions of cells. The step of cell multiplying does, but occurs with harmful cells. The old deviant cells carry on with their lives, and when they bite the dust, they ought to die, yet new ones are made in any event, when they are not required, which is the way the cancer develops [3]. Cellular breakdown in the lungs represents around one-fourth of all malignant growth fatalities and caused straight by cigarette smoking in 82% of examples, bringing about roughly 107,870 smoking-related cellular breakdown in the lungs passings in 2021 [4,5].Cellular breakdown in the lungs is liable for unbelievable 2.2 million new cases overall every year, or around 11.6% of all disease analyse. In Pakistan, cellular breakdown in the lungs is the second most regular malignant growth in men and the fifth most predominant in Woman. Cellular breakdown in the lungs remembered to cause 22,000 new cases and 19,000 fatalities in Pakistan every year. Tobacco use, environmental contaminants, and hereditary inclination all add to the nation's increasing cellular breakdown in the lungs rate. Pakistan is no exemption for this developing issue, with cellular breakdown in the lungs pervasiveness on the ascent, featuring the significance of bringing issues to light, early discovery, and giving available treatment choices[6].The annual rate of mortality for males with lung cancer decreased by around 5% between 2014 and 2020. For women, the annual death rates from LC(Lung cancer) decreased by 4% over the same time frame. According to research, these declines are due to fewer people starting to smoke, more people quitting, and breakthroughs in diagnosis and treatment. [7]

The Major Risk factors are below.

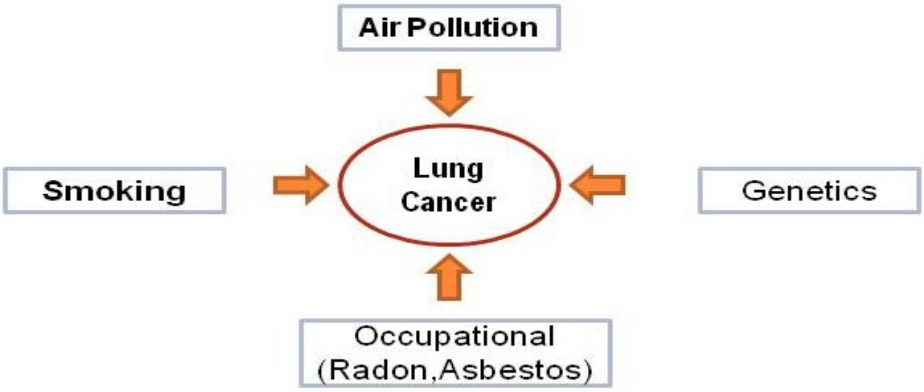


Figure 1. Major Factor effect on LC

Air contamination or pollution adversely effect human-health. In Pakistan, air contamination is a significant public health concern, and its impact on cellular breakdown in the lungs is turning out to be more serious. Vehicle emission, industrial activities, the burning of solid waste and harvest build-up, building activities, and fossil fuel energy production are significant supporters of air contamination in Pakistan. Urban communities with high air contamination levels, like Lahore, Karachi, and Islamabad, are especially impacted. While research has been directed on the effect of air contamination on health in Pakistan, its large portion is focused on case studies of major cities including Karachi, Lahore, Peshawar, Quetta, Rawalpindi, and Islamabad [8].According to research, the majority of people reside in areas where air pollution is a daily occurrence. Transportation and industrial emissions, the production of electricity, smoke from purposeful burning, and wildfires are common sources of pollution [9]. Numerous carcinogenic chemicals, including PAHs (Polycyclic aromatic hydrocarbons), benzene, and other heavy metals, are found in air pollution. When these pollutants enter the body through breathing, they can travel deep into the lungs and cause lung tissue damage. Significant air poisons, which are transmitted straightforwardly into the climate principally because of the burning of fossil and biomass fuels, incorporate vaporous toxins (like SO2, NO2, CO, and VOCs) and particulate matter (PM), which incorporates carbonaceous spray particles like dark sediment or soot [10]. Research has demonstrated that there is typically a greater incidence and mortality rate of lung cancer in places with higher air pollution levels.

Cellular breakdown of lung cells is mostly caused by smoking. Cigarettes contain about 7,000 compounds, at least 250 of which are toxic and more than 50 of which are known to cause cancer. These include things like formaldehyde, benzene, tar, and arsenic, among others. When tobacco is smoked, these compounds largely breathed into the lungs, although some of discharged into the environment. Numerous compounds included in tobacco smoke are carcinogens, which mean they can harm a cell's DNA. This harm may result in gene alterations that control cell division and proliferation, raising the possibility of malignant cells developing. Tobacco smoking is broadly archived as the absolute most significant modifiable risk-factor for cellular breakdown in the lungs [11,12]. As indicated by a meta investigation incorporating 287 examinations, smoking raised the risk of cellular breakdown in the lungs by more than five times [13].Inhaled tobacco smoke damages lung cells, leading to aberrant cell growth. Individuals who smoke intensely and additionally for a drawn out period have an expanded risk of creating disease. Regardless of whether an individual smoke, normal openness to smoke from another person's cigarettes, pipes or cigars can raise their possibility creating cellular breakdown in the lungs. We refer to this as "secondhand" or ETS (Environmental Tobacco Smoke)[9].

A positive family background of cellular breakdown in the lungs has been distinguished as a hazardous factor in various registry based examinations or studies, which have demonstrated a high familial risk for beginning stage LC[14]. Even after making thorough adjustments for smoking, increased relative risks were discovered [15]. Certain individuals are predisposed to cellular breakdown in the lungs genetically. Regardless of whether they not have habit of smoke cigarettes, the people who have an ancestors or sibling with cellular breakdown in the lungs may be more likely to get the disease themselves due to genetic abnormalities [9]. The genetic susceptibility to several complex diseases, including lung cancer, has successfully identified by GWAS [16]. [Skip 17]Furthermore, elevated DNA adduct levels have been connected to an expanded risk of LC, reflecting both openness to ecological cancer-causing agents and individual-vulnerability

Other pollutants, for example chemicals or gases encountered in environment or at working place ,can raise a individual's risk of processing lung cancer. People who cook with coal or wood fires worldwide have a increased threat of developing LC. Additionally, vapours from diesel, gas or melting metals may raise the risk of LC. Additional variables that could raise the hazard of lung cancer entail radiation exposure, nickel, chromium, nickel and arsenic-l. Inhaling asbestos such fibres can generate irritation in the lungs. Those People which are exposed with asbestos in industries for example, asbestos mining, shipbuilding, insulation and auto-mobile brake maintenance are more likely to get LC. Radon openness has been connected to an expanded risk of a few diseases, especially cellular breakdown in the lungs **[9].**

1.1. Objectives of the Study

The project aims to use advanced machine learning techniques to better comprehend the complex interaction between environmental variables and lung cancer. The following are the main objectives of this study:

Utilize machine-learning techniques to create predictive models that can foretell the hazard of lung cancer by examining a wide range of environmental factors in detail. To make hearty, versatile, and interpretable prescient models for cellular breakdown in the lungs risk, consolidate state of the art AI approaches including ANN, KNN, SVM, NB, D.T, RF, and LR. The primary goal is to foster prescient apparatuses that empower proof based clinical dynamic in cellular breakdown in the lungs treatment, empower early ID, give exact gamble characterization, and license altered avoidance mediations. Use AI calculations to distinguish and focus on natural variables influencing cellular breakdown in the lungs risk. This incorporates investigating communications between a few elements to track down complex examples that favor the beginning of cellular breakdown in the lungs, bringing about a more nuanced comprehension of their consolidated impacts.

In this research, I hope to provide a comprehensive and dependable ensemble learning framework that assemble a variety of machine learning models. The ultimate purpose of work is providing healthcare practitioners' and policymakers' access to reliable and generalizable forecasts by leveraging the ensemble's range.

2. Related Work

Early detection of lung cancer can help to save numerous lives. Lung disorders are conditions that affect the lungs, which are the organs involved in breathing. The paper **[18]** aims to identify and categorize lung disorders by utilizing moment invariants for efficient feature extraction, genetic algorithms for feature selection, NB and DT classifiers for final result classification. Preprocessing techniques are used to minimize noise in the image so that the vital information can be extracted. The data are trained, tested, and classified using the NB and DT classifiers. Compared to the naïve bayes classifier, the decision tree classifier produces more accurate results.

The paper [21] research that glances at the accuracy proportions of three different AI classifiers. SVM produced the best results, as demonstrated by the experiment. The precision for SVM was noted as 95.56%. Convolutional Neural Network (CNN) came about with 92.11% precision and kNN came about with of 88.40% exactness.

In the study [19] lung cancer is the world's greatest cause of mortality for people of all ages. It is particularly severe because it can be challenging to diagnose in its early stages. The study's informational indices came from UCI databases that contained patient information on lung cancer. SVM, KNN, and CNN were the three classifiers used in this study. This paper is based on the use of the WEKA Tool to investigate classification algorithm accuracy. The outcomes recommend that the SVM has the most elevated exactness, roughly 95.56 percent, and may distinguish cellular breakdown in the lungs in its beginning phases with a high precision proportion, saving multiple lives, but the K-Nearest Neighbor has a lower accuracy [25] [26].

This paper [20] stated that lung cancer remains a major issue worldwide. In this study, work propose a comprehensive technique that uses the power of several machine learning algorithms, such as Gradient Boost, Logistic Regression, LGBM, and Support Vector Machine, to address these difficulties and improve patient care. These algorithms were selected because they can effectively handle the complexity of LC data and provide precise case categorization and prediction. Using ML Models those are LR, K-NN, RF, GB, SVM and Light Gradient-Boosting Machine, Random Forest attained a stunning accuracy of almost 97%, demonstrating its capacity.

This study [23] [27] utilized picture acknowledgment and AI to make an assortment of PC supported frameworks. Different division, extraction of elements, characterization procedures like attentive wavelet change, Dark level co-event lattice, SVM, ANN. The DNN had ninety seven percent exactness, CNN had ninety four percent , ANN had ninety nine percent , and SVM ninety six percent, as per the analysts.

The paper [22] proposed a ANN for identifying regardless of whether cellular breakdown in the lungs is tracked down in the human body. Effects were utilized to analyze cellular breakdown in the lungs, these effects like Yellow fingers, Anxiety, Breath shortness, difficulty in swallowing, Sensitivity, and Wheezing etc. It utilized these and other data related the patient as information factors for the ANN Algorithm [24]. This model was prepared and approved utilizing the cellular breakdown in the lungs data points. The model that was proposed was looked at and tried. It provided 99.01 percent accuracy.

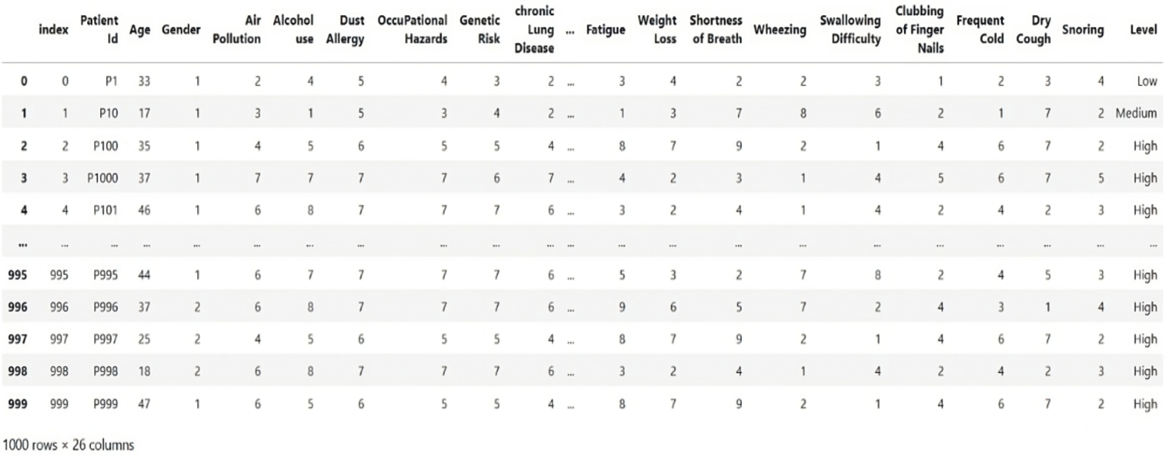
3. Material and Method

Lung cancer, similar to all cancers, can act in different ways in every person, subject on the type of lung cancer it is and the phase it is in. Nonetheless once lung cancer feasts outer the lungs, it frequently goes to similar places.

3.1. Data_Collection

Data set utilized for this examination work is gotten by the following site: [https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/code].

The dataset has 25 quality and 1000 occasion. An Identityfication number in the first column identifies every instance, and the level in the last column indicates whether the tumor has spread to a high, medium, or low level.



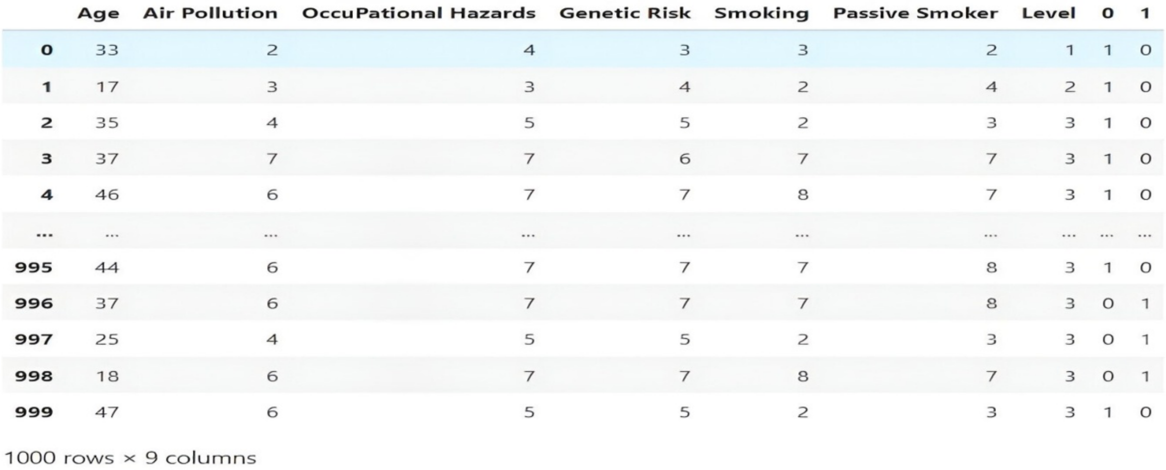
	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	1	2	3	4	Low
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	2	1	7	2	Medium
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	4	6	7	2	High
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	5	6	7	5	High
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	2	4	2	3	High
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	995	P995	44	1	6	7	7	7	7	6	...	5	3	2	7	8	2	4	5	3	High
996	996	P996	37	2	6	8	7	7	7	6	...	9	6	5	7	2	4	3	1	4	High
997	997	P997	25	2	4	5	6	5	5	4	...	8	7	9	2	1	4	6	7	2	High
998	998	P998	18	2	6	8	7	7	7	6	...	3	2	4	1	4	2	4	2	3	High
999	999	P999	47	1	6	5	6	5	5	4	...	8	7	9	2	1	4	6	7	2	High

1000 rows × 26 columns

Figure 2. LC Dataset

3.2 Data Extraction

By removing unnecessary Columns data contain the following columns also Data has the level of low, medium and high, these are replaced by 1,2,3 and make dummies so we get best prediction/Results.



	Age	Air Pollution	OccuPational Hazards	Genetic Risk	Smoking	Passive Smoker	Level	0	1
0	33	2	4	3	3	2	1	1	0
1	17	3	3	4	2	4	2	1	0
2	35	4	5	5	2	3	3	1	0
3	37	7	7	6	7	7	3	1	0
4	46	6	7	7	8	7	3	1	0
...	...	...	...	...	...	...	...	...	...
995	44	6	7	7	7	8	3	1	0
996	37	6	7	7	7	8	3	0	1
997	25	4	5	5	2	3	3	0	1
998	18	6	7	7	8	7	3	0	1
999	47	6	5	5	2	3	3	1	0

1000 rows × 9 columns

Figure 3. Feature exacted from Dataset

3.3. Classifications By Using Machine Learning Algorithms/Model Slection

3.3.1 Logistic Regression

A statistical technique for analyzing datasets in which an outcome is determined by one or more independent variables is known as logistic regression. Class is used to build this classification function, which makes use of a single estimator and a single multinomial logistic regression model. In a particular way, logistic regression usually says where the boundary between the classes is and that the class probabilities depend on how far away from the boundary they are. Most of the time, the result is binary (like true or false, 0/1, or yes by no). Based on one or more predictor variables, it makes predictions about the likelihood of a binary response.

3.3.2. Logistic Function:

- The logistic function is defined as: $\sigma_z = \frac{1}{1+e^{-z}}$,where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta n x_n$
- In logistic regression, the probability $P(y = 1/x)$ is given by:
- Log-Odds (Logit): $\log\left(\frac{P\left(y=\frac{1}{x}\right)}{1-P(y=1/x)}\right) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta n x_n)$
- Likelihood Function:
  $L(\beta)=\prod_{i=1}^{N} P(y_i|x_i;\beta)$
- For binary outcomes:
  $L(\beta)=\prod_{i=1}^{N} P(y_i = 1|x_i;\beta)\ \hat{}^{yi}\ .\ (1 - P(y_i = 1|x_i;\beta))^{1-yi}$

Logistic Regression is a powerful and interpretable classification algorithm. It models the probability of the binary outcome using a logistic function. The parameters are estimated by maximizing the log-likelihood function, often using gradient descent.

3.3.3. Naive Bayesian (NB) Networks:

Naive Bayes (NB) is a simple yet powerful probabilistic machine learning classifier based on Bayes' theorem with the assumption of independence between the features. It is particularly useful for classification tasks, such as spam detection, sentiment analysis, and text classification.

Bayes' Theorem: NB classifiers are based on the Bayes' theorem. The Bayes' theorem is expressed as:

$$\mathrm{P}\left(\frac{\mathrm{A}}{\mathrm{B}}\right) = \frac{\mathrm{P}\left(\frac{\mathrm{B}}{\mathrm{A}}\right)\cdot \mathrm{P}(\mathrm{A})}{\mathrm{P}(B)}$$

Where:

- $P(A/B)$ is a subsequent probability of class A given predictor B.
- $P(B/A)$ is the likelihood which is the probability of interpreter B given class A.
- $P(A)$ is a earlier probability of A.
- $P(B)$ is the prior probability of predictor B.

Naive Assumption: This assumption is that the features are independent of each other given the class.

$$P\left(\frac{B}{A}\right) = P(\{B1, B2, \ldots, Bn\}/A) = P(B_1/A)\cdot P(B_2/A)\cdot\ldots\cdot P(B_n/A)$$

Classification: Given a new instance to classify, the goal is to find the class A that maximizes the posterior *probability* $P(A/B)$. Using the independence assumption, this is calculated as:

Where:$\hat{y} = \underset{c\in\{c1,c2,c3\ldots ck\}}{\text{Max}} P(Ci)\prod_{j=i}^{n} P\langle xj|Ci\rangle$

- $\widehat{y}$ is the predicted class.
- C_i is a possible class.
- x_j is a feature

Prior Probability P(Level)): This is the fraction of instances that belong to each class in the training data.

$P(Level)$=$\frac{\text{Number of instances with Level}}{\text{Total number of instances}}$

These are very basic Bayesian networks made up of directed acyclic graphs with only one parent (the node that is not being observed) and several children (the nodes that are being observed), with a strong assumption that the child nodes are independent of their parent.

3.3.4. KNN Classifier

KNN is a classification and regression-based supervised machine learning algorithm. For classification and regression, the K-Nearest Neighbors (KNN) classifier is a straightforward, non-parametric, lazy learning algorithm. It gives a data point to the class that is most common among its k closest neighbors in classification. It was based on the straightforward idea that numbers from the same data set are likely to have labels that are similar.

Let's denote the dataset as $D = \{(x_0, y_0), (x_1, y_1), \ldots, (x_r, y_r)\}$, where x_i represents the independent variables and y_i represents the target variable (dependent variable, in this case, "Level").

Distance calculation(*Euclidean* Distance): $d(x, x_i) = \sqrt{(\sum(x_i - x_j)^2)}$

where x_i is the ith feature of the new instance x, and x_j is the jth feature of the ith instance in the training dataset.

k-nearest neighbors selection: $N_x = \{x_i \mid d(x, x_i) \leq d(x, x_k)\ for\ all\ x_i\ in\ D\}$

Where N_x is the set of k nearest neighbors to x, and x_k is the kth nearest neighbor.

Finding k-nn number: choose the k_number with the smallest distance with x.

Predict the target variable: If it's a classification problem, assign the new instance x the class label that occurs most frequently among its k nearest neighbors. If it's a regression problem, take the average of the target values of its k nearest neighbors.

When you have a new instance x without a label, you: $y_{pred} = mode\langle y_i | x_i\ in\ N_x\rangle$

where y_{pred} is the predicted target variable, and Mode is the most frequent value in the set of target variables of the k nearest neighbors.

3.3.5. Support Vector Machines (SVMs):

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. The goal of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

Hyperplane Equation:

$$w \cdot x + b = 0$$

Where $'w'$ is the weight vector, $'x'$ is the element vector and b is the predisposition term.

Classification Condition: For a data point x_i belonging to class $y_i \in \{-1,1\}$:

$$y_i(w.x_i + b) \geq 1$$

Objective Function: SVM aims to minimize the norm of the weight vector $\|w\|$ while satisfying the classification condition. This is equivalent to maximizing the margin.

The objective function for linear SVM is:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

Subject to :

$$y_i(w.x_i + b) \geq 1, \qquad \forall_i$$

Dual Problem: The optimization problem can be converted to its dual form using Lagrange multipliers:

$$\max_{a} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j (x_i x)$$

Subject to:

$$\sum_{i=1}^{N} a_i\, y_i = 0 \quad and \quad a_i \geq 0,\ \forall_i$$

Here $a_i$the Lagrange multipliers.

Kernel Trick:

SVM can map non-linearly separable data into a higher-dimensional space using the kernel function (x_i, y_j) , where a linear hyperplane can be used to separate the classes. Normal portions include:

Polynomial Kernel: $K(x_i, y_j) = (x_i . x_j + 1)^d$

Radial Basis Function (RBF) Kernel: $K(x_i, y_j) = e^{(-\gamma \|x_i - x_j\|^2)}$

Kernel Calculation:

Using a linear kernel, the kernel matrix K is computed as:

$$K(x_i, y_j) = x_i . y_j$$

Final Decision Boundary:

Once we have the optimal α values, the weight vector w can be computed:

$$w = \sum_{i=1}^{N} a_i y_i x_i$$

$$b = y_i - \sum_{j=1}^{N} a_j y_j \, K(x_j, x_i)$$

For the first support vector:

$$b = 1 - \sum_{i=1}^{N} a_j y_i \, K(x_j, x_1)$$

3.3.6. Decision Trees:

A supervised learning algorithm for classification and regression tasks is called a Decision Tree (DT). Decision Trees (DT) are trees that sort instances by their feature values to classify them. In a decision tree, each node represents a feature in an instance that needs to be classified, and each branch represents a possible value for the node. Starting at the root node, instances are categorized and sorted according to their feature values. When pruning decision trees using a validation set, decision tree classifiers typically use post-pruning methods to evaluate their performance. Any hub can be taken out and relegated the most widely recognized class of the preparation examples that are arranged to it [24]. It divides the data into subsets according to the value of the input features. A decision model resembling a tree is created by this recursive splitting.

Gini Impurity:

$$Gini(D) = \sum_{i=1}^{c} {p_i}^2$$

where p_i is the probability of class i in dataset D, and C is the total number of classes.

Information Gain:

$$IG(D, A) = Entropy(D) - \sum_{v \epsilon Values(A)} \frac{(D_v)}{D} . Entropy(D_v)$$

$$where \; Entropy(D) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Mean Squared Error (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Where y_i is the actual value and $\hat{y_i}$ is the predicted value.

Stopping Criteria:

Typical criteria for stopping include:

- The maximum tree depth
- The bare minimum of samples per leaf
- A reduction in the number of impurities

Final Tree Representation:

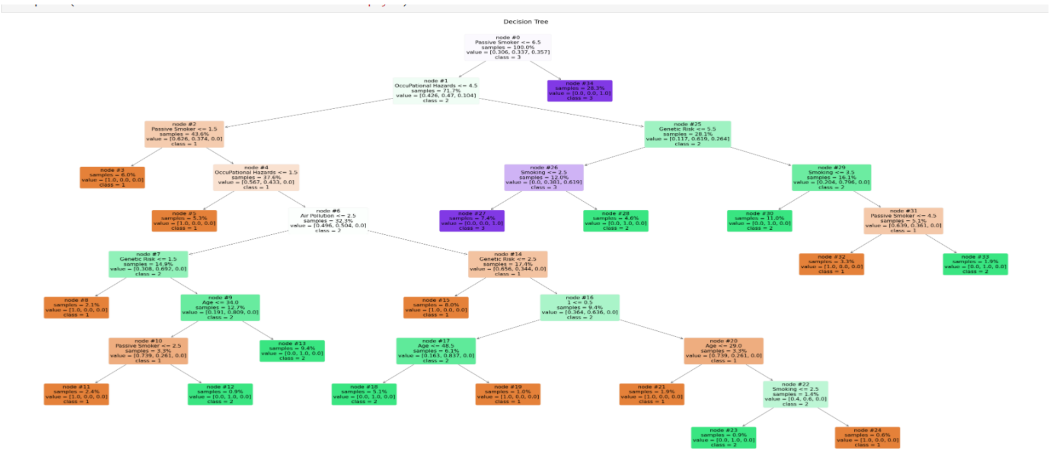


Figure 4. DT of LC Dataset

3.3.7. *Artificial Neural Networks (ANN):*

Artificial Neural Networks (ANN) are enlivened by the human brain and are utilized for errands like order, relapse, and example acknowledgment. They are made up of layers of neurons, and the weights that connect them are changed during training to reduce the prediction error.

Structure of ANN

1. The input features are received by the input layer.
2. Transforms and computations can be carried out in the hidden layers.
3. Creates the final result.

Each connection between neurons has an associated weight, and each neuron has an activation function that introduces non-linearity into the model.

Hidden Layer:

Let's say we have M neurons in a single hidden layer. A matrix called W_1 represents the weights that connect the input layer to the hidden layer, while a vector called b_1 represents the biases.

$$W_1 = \begin{pmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} & w_{16} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} & w_{26} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & w_{n4} & w_{n5} & w_{n6} \end{pmatrix}, b_1 = \begin{pmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{pmatrix}$$

Compute the linear combination z_1:

$$z_1 = W_1 \cdot x + b_1$$

Apply an activation function $f(\cdot)$, such as $ReLU$ (Rectified Linear Unit):

$$f(z_1) = \max(0, z_1)$$

Output Layer:

The hidden layer's output is connected to the output layer. Suppose the output layer has K neurons (for K classes). The weights and biases are W_{out} and b_{out}, respectively.

$$W_{out} = \begin{pmatrix} w_{out1,1} & w_{12} & w_{13} & w_{14} & \cdots & w_{out1M} \\ w_{out2,1} & w_{22} & w_{23} & w_{24} & \cdots & w_{out2,M} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{outn,1} & w_{outn2} & w_{outn3} & w_{outn4} & \cdots & w_{outnM} \end{pmatrix}, b_{out} = \begin{pmatrix} b_{out1} \\ b_{out2} \\ \vdots \\ b_{outn} \end{pmatrix}$$

$W_{out} = (0.1 \quad 0.2 \quad 0.3)$, $b_{out}(0.1)$

Compute the final output z_{out}:

$$z_{out} = W_{out} \cdot a_1 + b_{out}$$

Apply a suitable activation function $g(.)$, such as softmax for classification:

$$\hat{y} = g(z_{out}) = \frac{e^{z_{out}}}{\sum_{k=1}^{n} e^{z_{outk}}}$$

The predicted class $\hat{y}$ is:

$$\hat{y} = arg \max_{k} \hat{y}_k$$

Gradient Descent:

For weights in the hidden layer:

$$\frac{\partial L}{\partial W_1} = ((\hat{y} - y).\, w_{out}^T \circ \acute{f} z_1).\, x^T$$

Update weights using the learning rate η:

$$W_{out} \leftarrow W_{out} - \eta \frac{\partial L}{\partial W_{out}}$$

By applying these principles and detailed mathematical interpretations, ANNs can be effectively used to analyze complex datasets and derive meaningful insights from them.

3.3.8. Random Forest (RF)

RF is the gathering learning method utilized that for the arrangement, for the regression, and different undertakings. During training, multiple decision trees are constructed, and the mode of the classes (for classification) or mean prediction (for regression) of each tree is output.

Decision Tree Construction

For each bootstrap sample, construct a decision tree as follows:

Node Splitting Criteria: For classification and regression, make use of the Gini impurity or information gain and Mean Squared Error (MSE).

Gini Impurity for Classification:

Gini impurity for a node t is given by:

$$Gini(t) = 1 - \sum_{1}^{n} {p_i}^2$$

where p_i is the proportion of instances of class i in the node, and n is the number of classes.

For a split that divides data into subsets t_L and t_R, the Gini impurity for the split is:

$$Gini_{Split} = \frac{N_L}{N} Gini(t_L) + \frac{N_R}{N} Gini(t_R)$$

where N is the total number of instances, N_L and N_R are the number of instances in the left and right nodes, respectively.

MSE for Regression:

$$MSE(t) = \frac{1}{N_t} \sum_{i\epsilon t} (y_i - \hat{y}_t)^2$$

where N_t is the number of instances in node t, y_i is the true value, and $\hat{y}_t$ is the mean value of the target variable in node t.

3.3.9. ROC Curve

Mathematical Interpretation of ROC Curve

The ROC twist depends on two measures:

True positive Rate (TPR) or Mindfulness: Described as $TPR = \frac{TP}{TP+FN}$where TP is the amount of veritable up-sides and FN is the amount of deceiving negatives.

False Positive Rate (FPR): Described as $FPR = \frac{FP}{\mathrm{TN+FP}}$, where FP is the amount of positive sides and TN is the amount of veritable negatives.

3.3.10. Stacking

Computation stacking, generally called stacked hypothesis, is a gathering learning system that joins different base models to deal with the overall execution of the assumption. The idea is to set up a couple of base models and a short time later use a meta-model to join their assumptions. We will discuss the mathematical interpretation of stacking using base models such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), Artificial Neural Networks (ANN), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF).

Train each model:

$$M_{SVM}\ , M_{KNN}\ , M_{ANN}\ , M_{NB}, M_{DT}, M_{LR}, M_{RF}$$

Generate predictions for meta-training data:

$$D_{meta,train} = \{M_{SVM}(D_{train}), M_{KNN}(D_{train})\ , M_{ANN}(D_{train})\ , \\ M_{NB}(D_{train}), M_{DT}(D_{train}), M_{LR}(D_{train}), M_{RF}(D_{train})\}$$

Train the meta-model:

$$D_{meta}{=}train(D_{meta,train}, y_{train})$$
$$y_{pred} = M_{meta}(D_{meta,train})$$

By putting these models together in a stacking framework, we can take advantage of their individual strengths and possibly raise the overall accuracy of the predictions.

4. Results and Finding

The presentation of various ML models was evaluated and isolated considering their accuracy in expecting the objective variable. The models related with this affiliation were SVM, NB, (ANN), (KNN), (RF), (LR) and Decision Tree (DT).

The results, as depicted in the accuracy frame in the introduction of these models:

SVM achieved a precision of 68.67%, showing a moderate level of farsighted execution.

NB really squashed SVM with a precision of 69.67%, showing its ability as a probabilistic classifier notwithstanding its speculation of section an expected entryway.

ANN showed an immense improvement, achieving an accuracy of 88.33%. This highlights the strength of mind networks in getting jumbled models inside the data.

KNN showed a massive leap in execution with an exactness of 98.33%. This model's ability to coordinate events considering the closest planning models shows astoundingly strong.

RF and Decision Tree (DT) were the top performers, with exactnesses of 99.33% and 99.67% autonomously. The association strategy, which joins different decision trees, adds to its blasting show. Decision Tree, but scarcely higher, shows the power of tree-based models in supervising different data intricacies.

LR achieved an accuracy of 76.67%, showing its credibility as a straight model, particularly for twofold (for binary classification tasks) portrayal attempts.

The close to assessment incorporates that while standard models like SVM and NB give an activity, extra befuddling models like ANN, KNN, RF, and DT offer unmatched precision. The Random Forest and Decision Tree models, unequivocally, show close stunning precision, making them especially sensible for the farsighted occupation holding up be finished.

These disclosures suggest that for applications requiring high accuracy, tree-based models and neural networks are great. The encounters obtained from this assessment will ask the attestation concerning the most fitting model for sending in useful circumstances, ensuring ideal sensible execution.

Results of Used Algorithms are given below.

Table 1. Model Accuracies for Lung Cancer Dataset

Model	Accuracy	Precision	Recall	F1
RF	0.99	0.98	0.99	0.98
KNN	0.98	0.99	0.97	0.98
SVM	0.79	0.79	0.83	0.81
NB	60.66	0.54	0.57	0.55
Log_R	0.76	0.75	0.81	0.78
DT	0.99	1.00	0.99	0.99
ANN	0.94	0.95	0.99	0.97

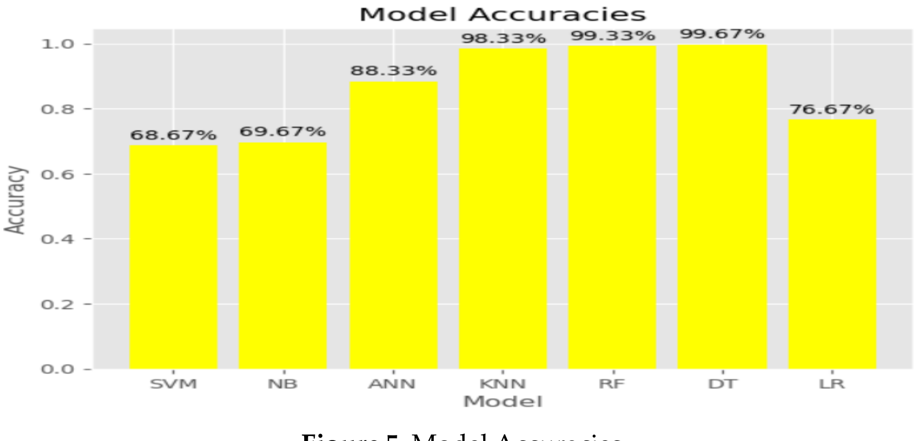


Figure 5. Model Accuracies

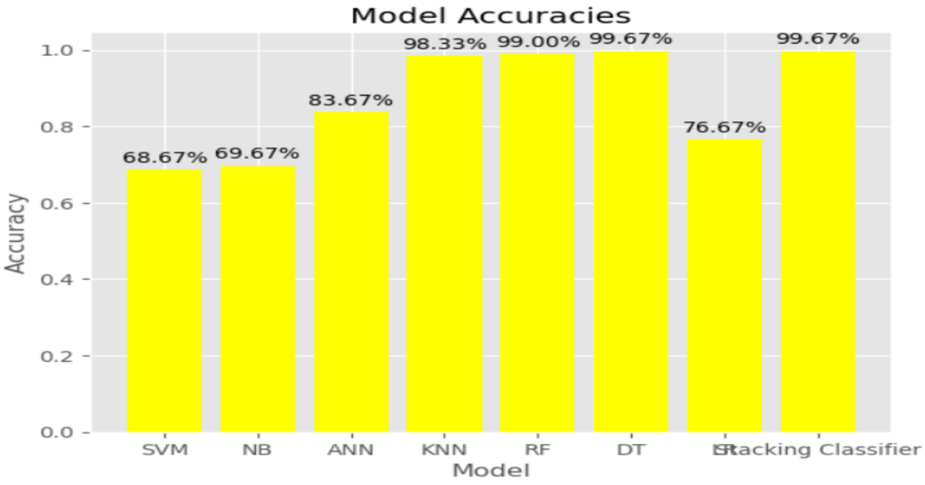


Figure 6. Stacking Model Accuracies

Using ROC Curve in Various Classifiers
Here is a layout of how ROC are made and utilized in various classifiers:

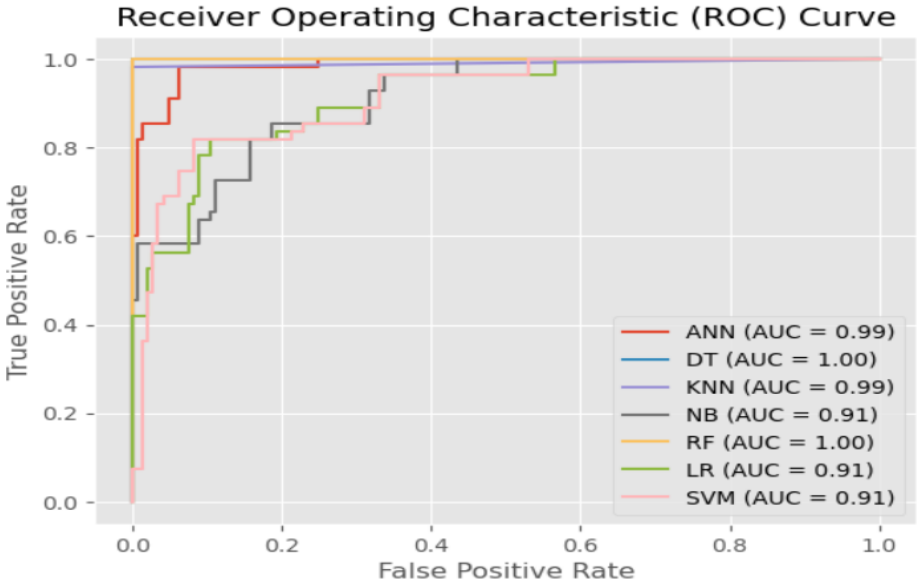


Figure 7. ROC for LC Dataset using ML Models

The ROC twist gives major areas of strength for a depiction to overview classifier execution across different edges. By taking a gander at the AUC (Locale Under the Curve), we can survey the overall limit of the classifiers to isolate between the classes over varied edges. This estimation is critical in practical settings like definite tests where changing responsiveness and explicitness is fundamental. Each classifier's method for managing resolving probabilities and scores gives obvious approaches to controlling edges and analyze the resulting ROC curve.

5. Future Direction

This section by uncovering knowledge into how different components add to the assumptions, this investigation can further develop straightforwardness and trust in the model, making it all the more alright to clinicians and other clinical consideration specialists. Additionally, research focused in on the interpretability of the social affair's assumptions is basic for understanding the major frameworks driving cell breakdown in the lungs results. This could incorporate making sensible man-made insight strategies and discernment contraptions to give clear and huge encounters from the model's expectations. At long last, watching out for moral examinations, for instance, mitigating tendencies, ensuring sensibility, and staying aware of data security, especially with fragile prosperity data, is basic for the trustworthy association of these models. Exploring mechanical advancements like quantum figuring and edge handling could similarly offer more compelling responses for complex artificial intelligence issues and update the speed of persistent assumptions. These broad future headings intend to develop serious solid areas for the spread out by the continuous revelations, ensuring unending improvement, practical congruity, and moral game plan of computer based intelligence models in understanding and expecting cell breakdown in the lungs results. By tending to these future headings, the exploration not just advances the specialized parts of cellular breakdown in the lungs expectation yet additionally guarantees that these headways convert into commonsense, moral, and successful clinical guide, improving by and large persistent consideration and wellbeing results. By coordinating computer based intelligence into cellular breakdown in the lungs expectation and the board, medical services suppliers can upgrade early location, work on analytic precision, and designer therapy systems to individual patients, eventually prompting better persistent results and more successful disease care.

More prominent Model Straightforwardness: Zeroing in on the interpretability of group expectations through reasonable simulated intelligence procedures and representation apparatuses will make it simpler for medical care experts to comprehend how expectations are made. This straightforwardness can assemble trust in the innovation, empowering its reception and guaranteeing that clinical choices depend on clear, justifiable information.

Moral and Fair computer based intelligence Sending: Tending to predispositions and guaranteeing reasonableness in prescient models will help in making evenhanded medical services arrangements. This implies that all patients, paying little mind to foundation, will get fair and impartial clinical evaluations, advancing inclusivity in medical services.

References

1. National Cancer Institute. Anatomy of the lung. Available at https://training.seer. cancer.gov/lung/anatomy, Accessed last time in January 2022.
2. American Cancer Society. What is lung cancer? Available at https://www.cancer. org/cancer/lung-cancer/about/what-is, Accessed last time in January 2022, 2019.
3. "What Is Cancer?" National Cancer Institute,www.cancer.gov/about- cancer/understanding/what-is-cancer.
4. Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. CA: A Cancer Journal for Clinicians, 71:7–33, 1 2021.
5. Quitting smoking after diagnosis of lung cancer improves survival and reduces the risk of disease progression. International Agency for Research on Cancer, Jul 2021.
6. https://idc.net.pk/lung-cancer-pakistan/
7. https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics
8. Anjum M.S., Ali S.M., Subhani M.A., Anwar M.N., Nizami A.S., Ashraf U. and Khokhar M. F. (2021). An emerged challenge of air pollution and ever-increasing particulate matter in Pakistan; a critical review. Journal of Hazardous Materials, 402, 123-943, DOI: 10.1016/j.jhazmat.2020.123943
9. https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/risk-factors-and-prevention, (2022)
10. Santos LR, Alves-Correia M, Camara M, et al. Multiple victims of carbon monoxide poisoning in the aftermath of a wildfire: a case series. Acta Med Port. 2018; 31: 146-151
11. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015
12. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017
13. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries
14. Bailey-Wilson JE, Sellers TA, Elston RC, et al. Evidence for a major gene effect in early-onset lung cancer. J La State Med Soc 1993; 145: 157–162.
15. 14 Lorenzo Bermejo J, Hemminki K. Familial lung cancer and aggregation of smoking habits: a simulation of the effect of shared environmental factors on the familial risk of cancer. Cancer Epidemiol Biomarkers Prev 2005; 14: 1738–1740.
16. Bossé Y, Amos CI. A decade of GWAS results in lung cancer. Cancer Epidemiol Biomarkers Prev 2018;27:363–379.
17. Bhuvaneswari, C., Aruna, P., & Loganathan, D. (2014). Classification of lung diseases by image processing techniques using computed tomography images. International Journal of Advanced Computer Research, 4(1), 87.
18. Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Academic Journal, 1(2), 141-149.
19. Kubra Tuncal, Boran Sekeroglu, and Cagri Ozkan, " Lung Cancer Incidence Prediction Using Machine Learning Algorithms", Journal of Advances in Information Technology Vol. 11, No. 2, May 2020
20. D. M. Abdullah, A.M. Abdulazeez and A.B. Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques", Qubahan Academic Journal, Vol. 1, no. 2, pp. 141-149, 2021, doi: 10.48161.
21. O. Mohammed et al., "Artificial Neural Network for Lung Cancer Detection," vol. 4, no. 11, pp. 1–7, 2020.
22. S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions," J. Imaging, vol. 6, no. 12, p. 131, 2020, doi: 10.3390/jimaging6120131.
23. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007). Pp. 249 – 268. Retrieved from IJS website: http://wen.ijs.si/ojs-2.4.3/index.php/informatica/article/download/148/140.
24. Khan, M. I., Imran, A., Butt, A. H., & Butt, A. U. R. (2021). Activity detection of elderly people using smartphone accelerometer and machine learning methods. International Journal of Innovations in Science & Technology, 3(4), 186-197. 50sea.
25. Ejaz, F., Ahmad, A., & Hanif, K. (2020). Prevalence of diabetic foot ulcer in lahore, Pakistan: a cross sectional study. Asian Journal of Allied Health Sciences (AJAHS), 34-38.

26. Khan, I., Siddique, M. Z., Butt, A. U. R., Mudassir, A. I., Qadir, M. A., & Munir, S. (2021). Towards skin cancer classification using machine learning and deep learning algorithms: A comparison. International Journal of Innovations in Science & Technology, 3, 110-118. 50sea.
27. Khan, M. F., Iftikhar, A., Anwar, H., & Ramay, S. A. (2024). Brain Tumor Segmentation and Classification using Optimized Deep Learning. Journal of Computing & Biomedical Informatics, 7(01), 632-640.