

Artificial Intelligence in Smoking Residue Detection: Bridging the Gap Between Medical Diagnostics and Predictive Analysis

A. Maiti, A. Roy, C. Dutta, and D. Saha

Abstract-- The objective of this work was to create a model that could identify smoking traces in the body and forecast future smoking propensity using a variety of health-related variables. Effective detection and monitoring of smoking residues in people is essential for identifying smoking behaviors and evaluating health concerns. The researchers used a cutting-edge strategy that combined medical diagnostics with artificial intelligence (AI) to enable advanced detection of smoking residues in order to overcome this barrier. The suggested methodology makes use of medical diagnostic tools, including an individual's lipid profile and dental test, to record and examine physiological and chemical indications connected to smoking. The vast data generated by modern medical diagnostic methods are meticulously analyzed and comprehended by AI-based systems to get improved accuracy and effectiveness of detecting smoking residue. Voluminous data sets serve as a crucial training ground for machine learning models, enabling them to discern patterns and accurately classify individuals based on their smoking habits. The study demonstrated a 99% prediction performance, making it a valuable tool for healthcare institutions to better understand and predict the likelihood of hospital admissions related to smoking. In the future, the study aims to determine the concentration of nicotine or cotinine and detect heart disease and lung conditions.

Index Terms--Evidence of Smoking; Medical Tests; Blood Pressure Test; Lipid Profile; Methodology

I. INTRODUCTION

Artificial intelligence was simply an idea not long back, but thanks to its remarkable pace, it has now permeated practically every aspect of modern life. The medical industry, driverless automobiles, smart homes, and other industries have all been transformed by artificial intelligence (AI).

In these areas, its achievements have considerably improved efficiency and dependability. Our research aims to identify smoking-related indications in the human body while utilising medical diagnostic techniques. We've delineated the medical criteria that promise dependable outcomes. Globally, tobacco use among the young population has reached alarming proportions.

There's a substantial body of evidence supporting the adverse impacts of tobacco consumption, including smoking, on chronic illnesses, which constitute 75% of healthcare expenditures in the United States. Despite progress since the initial Surgeon General's report on smoking and health in 1969, over a quarter of high school seniors persist in regular smoking [1]. Most adolescent smokers transition into habitual adult smokers. Remarkably, cigarettes, a legitimate consumer product, lead to premature death in half of their long-term users. This epidemic continues to propagate in low- and middle-income countries, ill-equipped to handle the ensuing health and economic fallout, whilst it continues to claim lives in the United States. Youth are often introduced to tobacco as a precursor to illicit drug use, which it precedes and makes more likely. Community-based interventions have proven effective in curbing youth tobacco use, yet cigarette marketing and promotional activities still pose a significant risk, enticing young individuals to initiate smoking. These findings, as detailed in the current study, remain relevant, accurate, and applicable [2]. However, since 1994, a substantial volume of research has been conducted that has substantially deepened our comprehension of the dynamics surrounding adolescent tobacco use, its prevention, and cessation. Therefore, the necessity for the current report is more urgent than ever.

II. ARTIFICIAL INTELLIGENCE TERMINOLOGIES

Artificial intelligence is a simulation of human intelligence executed by computer systems. Some intuitive tasks such as reasoning, knowledge representation, learning, NLP, perception, and manipulation of physical surroundings come under this term [2]. A subset of "artificial intelligence" is "machine learning", where the computer automatically learns the patterns of a task and improves the system to perform the task. There are two methods to

A. Maiti, MCA, Assistant Professor, Guru Nanak Institute of Technology, Kolkata, (e-mail: ananjan.maiti@gmail.com).

A. Roy, MCA, Guru Nanak Institute of Technology, Kolkata (e-mail: iamarijeet1999@gmail.com).

C. Dutta, MCA, Assistant Professor, Guru Nanak Institute of Technology, Kolkata (e-mail: chiranjib.dutta@gnit.ac.in).

D. Saha, MCA, Assistant Professor, Guru Nanak Institute of Technology, Kolkata (e-mail: dola.saha@gnit.ac.in).

learn, “supervised learning” and “unsupervised learning”. In case of “supervised learning”, the computer infers a function that delivers the input with the given parameters. Whereas “unsupervised learning” discovers a pattern automatically [3]. In this procedure, statistical models are fitted (or “trained”) in order to identify patterns in the data and predict future data with a similar distribution. Some popular models include “support vector machines (SVM)”, “random forest classifiers”, “logistic regression”, and “decision tree classifiers”.

III. OBJECTIVE OF THE STUDY

This study aims to engineer a machine-learning model capable of precisely estimating the probability of detecting smoking residue within an individual. The model leverages multiple health indicators, including fasting blood glucose, cholesterol, and blood pressure, dental hygiene, and lipid profiles, to anticipate traces of smoking. The objective is to provide a dependable and precise predictive framework to aid healthcare practitioners in identifying patients at increased risk of smoking, thus informing suitable interventions. The potential of this model to act as a valuable instrument for healthcare establishments in predicting the likelihood of smoking-related hospital admissions is underscored. Additionally, the study recommends future explorations into discerning the concentration of nicotine or cotinine and evaluating lung health.

The text in focus relates to the creation of a machine-learning model for detecting signs of smoking within an individual. Here are five potential applications of this topic:

- I. Healthcare institutions could leverage this model to identify patients at an elevated risk of smoking, guiding proper treatment.
- II. A significant opportunity to increase accessibility for medical practitioners is provided by the model's integration with current “electronic health record (EHR) systems”. By integrating the model with these systems, healthcare providers will have better access to information, which will improve the effectiveness and efficiency of their work.
- III. This model can be employed to spot smoking residue in the body, providing a comprehensive view of an individual's health status, including cholesterol, blood glucose levels, blood pressure, dental health, and lipid profiles.
- IV. Prospective enhancements to the model could include determining the concentration of

nicotine or cotinine present in the body and detecting cardiac and pulmonary conditions.

IV. EVIDENCE OF SMOKING

As we know the traces of smoking are nothing but the traces of nicotine that stays when a person is exposed to smoking (or using tobacco). With each exposure to smoking, we carry marks of nicotine with us [4]. Generally, the traces of nicotine stays within our body for 1 to 3 days after we quit smoking. Smoking does only leave traces of nicotine but also a substance called cotinine which stays for around 10 days after we quit smoking [5]. Parameters that play a crucial part in making our project a great success are: “systolic, relaxation, fasting blood sugar, cholesterol, triglyceride, LDL & HDL cholesterol, hemoglobin, urine protein, serum creatinine, AST, ALT, GTP, oral examination status, dental caries, and tartar status” [6].

V. MEDICAL TESTS

A. Test of Blood Pressure

Blood pressure, relaxation blood pressure and, systolic blood pressure levels can be very helpful to determine one's health. A healthy blood pressure level can determine a healthy body otherwise measures need to be taken to improve health and numbers [7]. We do not see any vast difference in the measurements of the systolic blood pressure, maybe some acute difference but a significant increase in diastolic blood pressure can be seen among smokers than non-smokers [8].

1) Blood Pressure (Systolic)

Every time our heart beats, the blood from the heart exerts some force on the walls of the arteries. The force exerted on the wall of the arteries is measured in the blood pressure measurement, called systolic blood pressure [9]. The blood pressure measurement contains two values, an upper value, and a lower value. The upper value is called systolic blood pressure level.

2) Blood Pressure (Diastolic)

The artery walls also experience force on them in between every beat of the heart. This force is also measured while measuring blood pressure, called diastole blood pressure [10]. This is the lower measurement value of the blood pressure calculation. In our project, we are referring to diastole as relaxation.

B. Hemoglobin Test

Hemoglobin is a type of protein that is found in the blood. More specifically, it is found in the red blood cells of the blood. Hemoglobin carries oxygen to the other parts of our body directly from the lungs [11]. The

level of hemoglobin increases in the blood if one smokes. Increased levels of hemoglobin may not be the only cause of smoking, due to other circumstances the levels of hemoglobin may increase.

C. Urine Protein Test

The test measures the presence of all the proteins and their ratio or quantities in the urine. It also specifies the presence of albumin in the urine. Albumin measurements are not required for the detection. In a normal case scenario, our urine does not have any protein. Our kidneys extract the proteins and stop them from flowing out of our bodies [12]. If protein can be found in urine indicates the stress on kidneys and cannot work properly, which indicates that the person may be exposed to smoking. The malfunctioning of kidneys may also have other prospects and smoking is one of them.

D. Serum Creatinine Test

Kidneys filter the wastes from our blood and eject them from our bodies with the help of urination. The creatinine test would provide how the kidneys are working, and how well or badly the kidney can perform the filtration process [13]. The levels of serum creatinine tend to decrease in people who are currently exposed to smoking, whereas the levels of serum cystatin C measurements does not show any significant difference.

E. Lipid Profile

1) Triglyceride

A type of fat or *lipid* that is found in our blood is called triglyceride. The levels of triglycerides are also important to keep watch. An increased level of triglyceride indicates probable heart disease [14]. Triglyceride measurements are found in the lipid profile. The levels of triglyceride are found to rise among smokers more than non-smokers. As a result, smoking increases the chance of heart disease, and a rise in triglyceride is one of the reasons or indicators.

2) Low density Lipoprotein

LDL represents “low-density lipoprotein”. This is called the “bad cholesterol” that can be found in our blood. This makes up most of the cholesterol in our body. The rise in the LDL measurements depicts that the bad cholesterol is increasing, which is not good for our hearts [15]. The rise in LDL levels may cause heart disease. The level of LDL is higher in the case of smokers than that of non-smokers. Cholesterol is ‘stickier’ as we arise exposed to smoking, so it clings to our artery walls and gets clogged.

3) High density Lipoprotein

HDL refers to “high-density lipoprotein”. HDL is the exact opposite of LDL. This is the good cholesterol that is found in our bodies [16]. The level of HDL decreases for smokers which is not good for our health, specifically for our hearts. This is the adverse effect of smoking.

F. Other Tests

“AST, ALT, GTP, oral examination status, dental caries, and tartar status” are also some parameters that are adversely affected due to smoking. Each test is also requested is proceed further with the test to check the body’s traces of smoking.

VI. METHODOLOGY

The suggested method's major goal is to forecast the likelihood of the presence of traces of smoking quickly and accurately diagnose the condition. We have considered several “machine learning algorithms”, including “Naive Bayes”, “k Nearest Neighbours (KNN)”, “Decision tree”, “Artificial Neural Network (ANN)”, and “Random Forest”, that can predict the traces of smoking depending on a few medical variables.

A. Data Collection

We require a dataset with details on the health, lifestyle, and demographics of individuals to develop a “heart disease prediction model”. The goal variable, which is the existence or absence of cardiac disease, should also be included in the dataset. The Framingham Heart Study, the Cleveland Heart Disease Dataset, and the Hungarian Institute of Cardiology Heart Disease Dataset are only a few of the publicly accessible datasets for heart disease prediction [17]. To create a more reliable model, we can select any one of these datasets or combine them.

B. Data Pre-Processing

Preprocessing the dataset is important before we train our “machine learning model”. Before we can decide how to proceed, we must first check the dataset for missing values. Categorical variables encoding certain components of the dataset may be categorical, such as gender and smoking status. These category variables must be converted into numerical values before being used in our machine-learning model [18]. We might need to scale the features to ensure that they all have the same range and distribution across the dataset.

C. Dataset Attributes

TABLE I
The Dataset attributes

	Attributes	Description	Values
1	Gender	Gender of Patient	M=Male F=Female
2	Age	Patient's age in years	Continuous Value
3	Height(cm)	Patient's height in Cm	Continuous Value
4	Weight(kg)	Patient's weight in Kg	Continuous Value
5	Waist(cm)	Patient's waist in Cm	Continuous Value
6	Eyesight(left)	Patient's eyesight in Left	range between 0.1 to 9.9
7	Eyesight(right)	Patient's eyesight in Right	range between 0.1 to 9.9
8	Hearing(left)	Patient's hearing in Left	1= Pass 2= Fail
9	Hearing(right)	Patient's hearing in Right	1= Pass 2= Fail
10	Systolic	The upper value of patient's BP	Continuous Value
11	Relaxation	The lower measurement value of the blood pressure calculation	Continuous Value
12	Fasting blood sugar	suger level in blood	Continuous Value
13	Cholesterol	cholesterol of Patient	Continuous Value
14	Triglyceride	triglyceride of Patient	Continuous Value
15	HDL	good cholesterol that is found in our body	Continuous Value
16	LDL	the bad cholesterol that is found in our blood	Continuous Value
17	Hemoglobin	hemoglobin of Patient	Continuous Value
18	Urine protein	Urine protein stands for alanine transaminase. It is an enzyme found mostly in the liver	Continuous Value
19	Serum creatinine	Serum creatinine stands for alanine transaminase. It is an enzyme found mostly in the liver	Continuous Value
20	AST	AST stands for alanine transaminase. It is an enzyme found mostly in the liver	Continuous Value
21	ALT	ALT stands for alanine transaminase. It is an enzyme found mostly in the liver	Continuous Value
22	GTP	GTP stands for alanine transaminase. It is an enzyme found mostly in the liver	Continuous Value
23	Oral	oral of Patient	Y= Yes
24	Dental Caries	dental caries stands for alanine transaminase. It is an enzyme found mostly in the liver	0= No 1=Yes
25	Tartar	Tartar of Patient	Y= Yes N=No
26	Smoking	detecting the traces of smoking within the body	Traces Found Traces not Found

The dataset has been collected from kaggle [19]. The description of the attributes of the dataset is elaborated briefly down below. The dataset employed in this study includes various attributes related to smoking's physiological impact:

- I. Hearing: Many research studies have emphasized the adverse effect smoking has on auditory capacities.
- II. Relaxation: This denotes diastolic blood pressure, also measured in "mmHg."
- III. Systolic: This represents systolic blood pressure, measured in "mmHg."
- IV. Cholesterol: The impact of smoking often manifests in elevated cholesterol levels. It's measured in "mg/dL."
- V. Triglyceride: Similarly, smoking leads to heightened triglyceride levels, measured in "mmol/L."
- VI. Fasting Blood Sugar: Smokers typically exhibit elevated blood sugar levels. This attribute is measured in "mmol/L or mg/dL."
- VII. LDL: Another adverse effect of smoking is increased LDL levels, measured in "mmol/L" or "mg/dL."
- VIII. HDL: Contrarily, HDL levels are observed to decrease due to smoking. The unit of measurement is "mmol/L" or "mg/dL."
- IX. Urine Protein: Protein presence in a smoker's urine is another attribute, measured in "mg/dL."

- X. Hemoglobin: Smokers usually have higher hemoglobin levels than non-smokers. This is measured in "g/dL."
- XI. Serum Creatinine: Among smokers, serum creatinine levels are typically high, measured in "mg/dL."
- XII. Dental caries: Representing tooth decay prevalent among smokers, its presence can be determined as either a yes or no.
- XIII. AST, ALT & GGT: Noted as GTP in the dataset, the levels of these attributes increase with daily smoking. They're measured in "IU/L."
- XIV. Tartar: This is the tar residue from smoking. Its presence or absence is noted in the dataset.

D. Features Selection

Before training our machine learning model, we must select the features from the dataset that are most important. Simplifying the model by feature selection can improve model performance. We can use techniques like "correlation analysis", "principal component analysis", or "recursive feature elimination" to ascertain which features are most important.

E. Model Training

We used train-test-split at first and used our models on them under iterations. The dataset has been split into separate training and testing datasets after numerous iterations. The train and test datasets that produce the best accuracy have been determined and saved for further use.

F. Methodology

We have directly imported the "train and test datasets" into "x-train, x-test, and y-train, y-test". Then OneHotEncoder is used over "x-train and x-test" to automatically convert the string attributes into integer datatype. Further, we put it into ColumnTransformer to reshape the data that will be passing through. Then simply we make a pipeline that consists of the instance of the model we are using and the ColumnTransformer. Due to ColumnTransformer & OneHotEncoder, we do not have to manually do the change of datatypes and the reshape of input data, which makes the model faster, i.e., less time-consuming.

G. Data Visualization

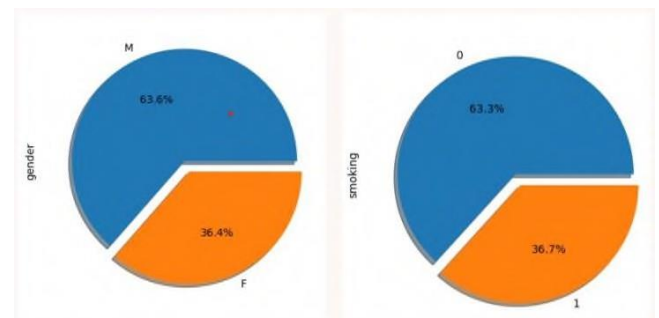


Fig. 1. smoking habits according to gender and overall status.

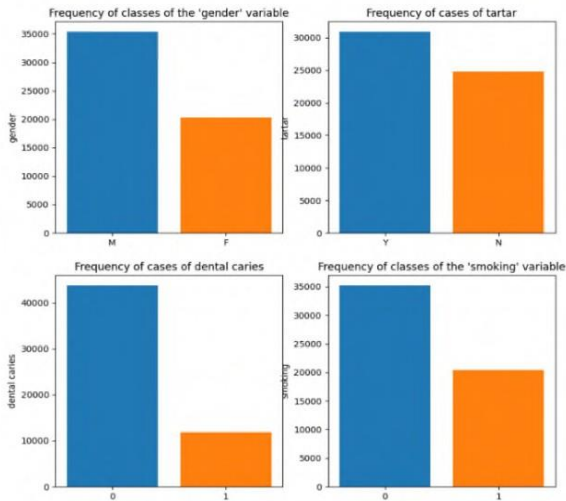


Fig. 2. A bar plot depicting the incidence of tartar, dental caries, smoking, and gender among both males and females.

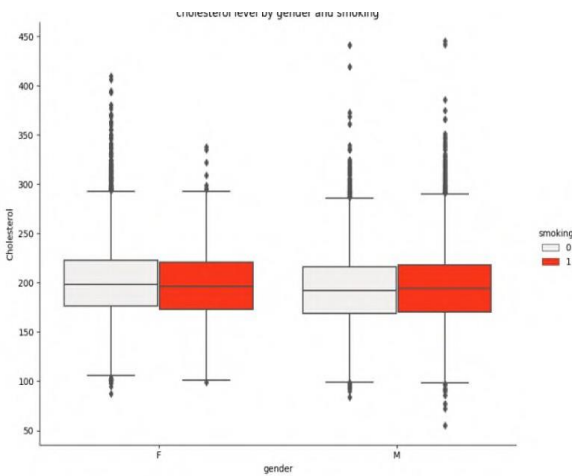


Fig. 3. Cholesterol by Gender CATPLOT

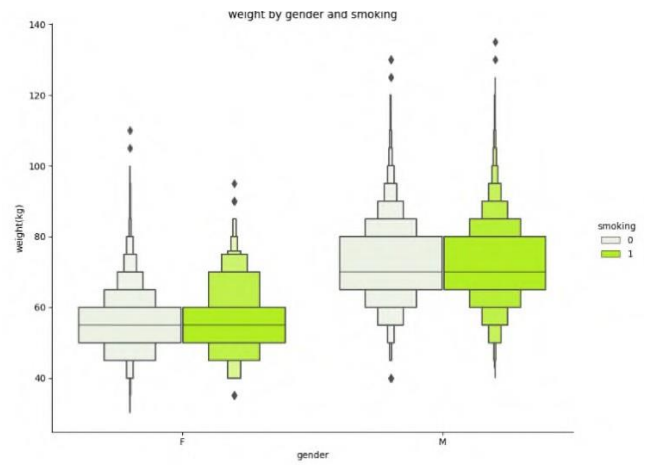


Fig. 4. Weight by Gender CATPLOT

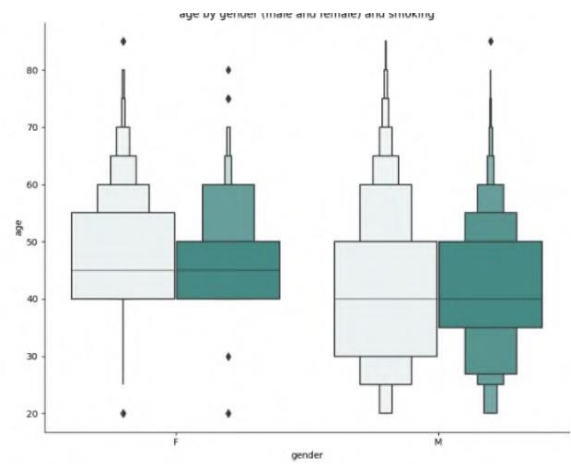


Fig. 5. Age by gender CATPLOT

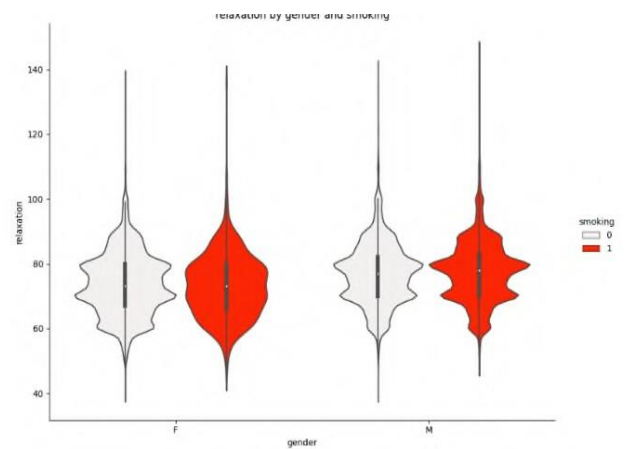


Fig. 6. Relaxation by gender CATPLOT

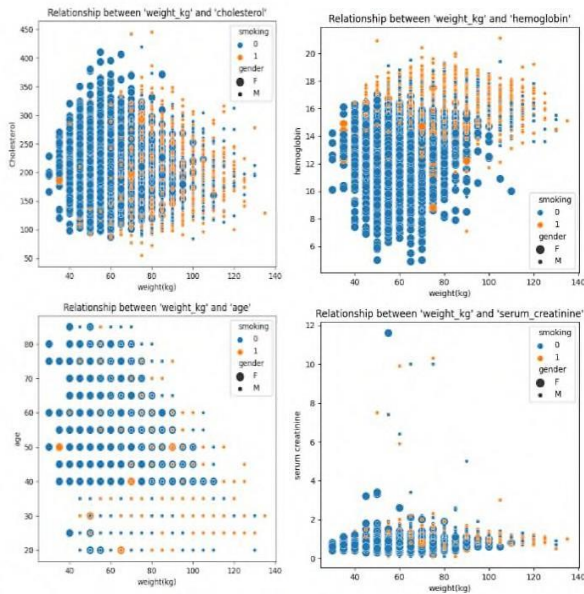


Fig. 7. Here is a scatter plot that displays the relationship between the data points.

H. Model Selection and Training

Following data pre-processing and feature selection, we can initiate the training process in our machine-learning model. A variety of methods, encompassing "logistic regression, support vector machines, decision trees, and neural networks," can be employed to predict smoking traces. Utilizing our dataset, we have the capability to assess the efficacy of disparate algorithms and pinpoint the most effective one. Moreover, it's crucial to partition the dataset into training and testing subsets to gauge our model's performance accurately. Methods such as "k-fold cross-validation" may be applied to enhance the assessment of our model. The selection of the model ought to be premised on achieving the highest degree of accuracy.

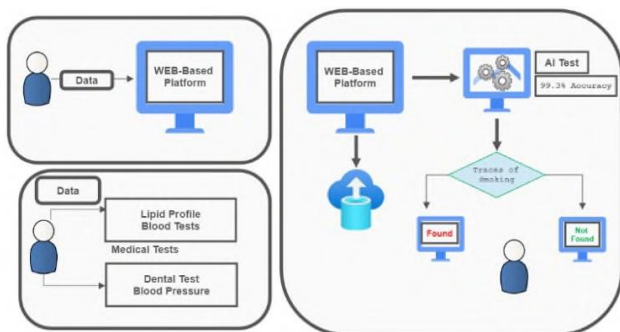


Fig. 8. Figure showing the workflow of our project

I. Deployment

Our machine-learning model can be deployed in a real scenario after developing and evaluating it. We can design a web-based application or integrate it with an existing electronic health record system to make our model easily accessible to healthcare professionals. In conclusion, while developing a “machine learning model for heart disease prediction”, careful data preparation, feature selection, algorithm selection, and model assessment are required. Healthcare experts have the ability to spot those who are more likely to develop heart disease early on, allowing them to provide prompt and effective treatment. This preventive strategy has the potential to improve general health outcomes and save lives. An accurate and reliable prediction framework enables it as well.

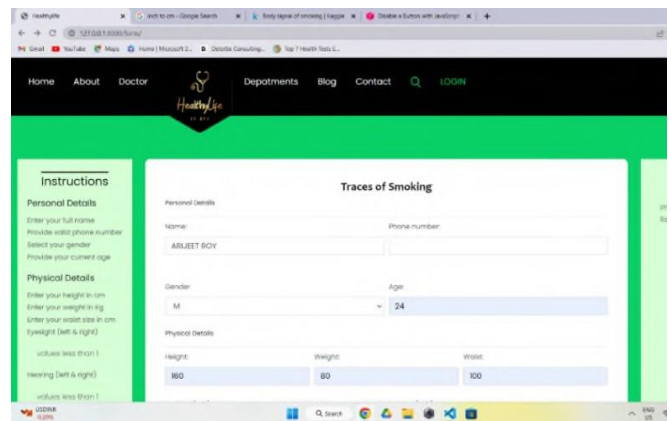


Fig. 9. Here is the template for the webpage.

VI. MACHINE LEARNING ALGORITHM

A. Logistic Regression

An analytical statistical model called a logit model is frequently used in predictive analytics and categorisation. This kind of model, also known as “logistic regression”, use predictive analytics to ascertain the probability that an event will occur. This is accomplished by using a number of independent variables, such as whether a person votes or not. In logistic regression, the dependent variable is a probability with a range of 0 to 1. “Logistic regression” adjusts the data to produce precise predictions by using the logit formula, which modifies the odds (i.e., “the probability of success divided by the probability of failure”). This logistic function, often known as the log odds or the natural logarithm of odds, is represented by the following formulas:

$$\text{Logit}(\pi) = 1/(1+\exp(-\pi))$$

$$\text{Ln}(\pi/(1-\pi)) = \text{Beta}_0 + \text{Beta}_1 * X_1 + \dots + \text{B}_k * K_k$$



The dependent or response variable in the “logistic regression equation” is marked by “Logit (π)”, whereas the letter x indicates the independent variable. In this model, the estimate of the beta parameter or coefficient is often accomplished using the well-known technique referred to as “maximum likelihood estimation (MLE).” We use MLE to iteratively evaluate various beta values to find the log odds’ best match. The goal of “logistic regression” is to maximise this function to get the highest accurate parameter estimate. A log-likelihood function is created after each iteration. We can identify the best-fitting model that faithfully captures the connection between the variables thanks to this iterative procedure. The forecast probability can be created by computing the conditional probabilities for each observation, taking their logarithm, and adding them all together after the optimal coefficient (or coefficients, in the case of several independent variables) has been identified. By using this method, we can get a trustworthy estimation of the likelihood or probability connected to the provided observations. We may successfully create a forecast probability that reflects the link between the variables under investigation by combining these logged conditional probabilities. This method aids in developing predictions that are based on the observed data and the discovered coefficients. In case of binary categorisation, a probability of less than 0.5 implies the prediction to 0 and a probability values more than 0.5 implies 1. The “Hosmer-Lemeshow test” is a well-liked technique for evaluating model fit.

There are three different kinds of categorical response-based logistic regression models.

1) Logistic regression (Binary)

There are only two possible outcomes when using this strategy because the response or dependent variable is binary (for example, 0 or 1). It is frequently used to determine if an email is spam or not, as well as whether tumors are malignant or not [20]. The technique frequently used in “logistic regression” is very common and one of the most widely used classifiers for binary classification tasks in numerous areas.

2) Logistic regression (Multinomial)

The dependent variable in the specific “logistic regression model” under discussion has three or more possible values but no established hierarchy between them. An illustration of such a scenario is when film studios try to foretell the kind of movie a spectator is likely to see in order to better market their films. The studio may use a “multinomial logistic regression model” to determine how much of an effect a person’s age, gender, and dating status may have on the type of films they love [20]. Thus, the studio may focus the advertising for a certain film on the demographic most probability to go see it.

3) Logistic regression (Ordinal)

“Logistic regression” is typically employed when the dependent variable comprises three or more distinct values organized in a specific sequence. This methodology is especially apt in scenarios featuring ordinal responses, such as letter grades ranging from A to F, or numerical rating scales from 1 to 5.

Positioned within the realm of “supervised machine learning” in “artificial intelligence,” “logistic regression” falls under the category of discriminative models. Its primary objective is the differentiation between multiple classes [20]. However, unlike generative algorithms like “naive Bayes,” “logistic regression” does not yield specific data about the class it attempts to predict—for instance, an image of a cat. From a “machine learning” perspective, there’s a nuanced adjustment in how the “beta coefficients” in “logistic regression” are computed. A loss function, specifically the negative log-likelihood, in tandem with the “gradient descent algorithm” is utilized to find the global maximum. This approach can derive estimates consistent with earlier ones.

Furthermore, “logistic regression” is prone to overfitting, particularly when the model is populated with an excessive number of predictive variables [20]. When dealing with high dimensional models, regularization techniques are typically employed, which levy penalties on sizable coefficients in the model’s parameters. This enhances the model’s ability to generalize and averts overfitting. The “Scikit-learn” library is a valuable resource for those wishing to delve deeper into the “logistic regression machine learning model,” as it provides comprehensive insights.

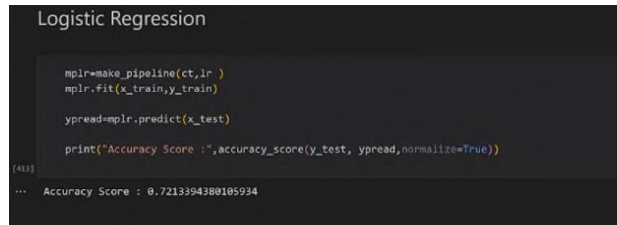
Logistic regression is widely applied for categorization and predictive tasks. Notable application scenarios include:

Fraud detection: “Logistic regression models” can serve as potent instruments in identifying data irregularities indicative of fraudulent activities. Financial institutions commonly identify specific behaviors or characteristics often associated with fraudulent transactions to bolster customer security. Additionally, SaaS-based companies have adopted these procedures to detect and eliminate counterfeit user identities from their datasets during data analysis for business performance assessment. By employing “logistic regression models,” these teams can effectively identify potential instances of fraud, enhancing security measures, and ensuring a more precise data analysis.

Disease Prediction: This analytics approach can also be utilised in the realm of medicine to forecast the likelihood that a given group would contract a particular disease or condition. Healthcare organisations can proactively build preventative care measures for people who are more likely to contract a specific illness by utilising predictive analytics. With this proactive approach, possible health risks can be reduced and general wellbeing can be promoted by early interventions and tailored healthcare interventions.

Churn Prediction: Different organizational functions may exhibit particular behaviors that are indicative of churn. For instance, management and human resources teams may want to know whether there are high performers inside the

firm who may leave; this kind of information can spark discussions to address problem areas within the company, such as culture or salary. Instead, the sales team would prefer to know which of its clients might choose to conduct business elsewhere. Teams might be inspired by this to develop a retention strategy to prevent revenue loss.



```

Logistic Regression

mplr=make_pipeline(ct,lr )
mplr.fit(x_train,y_train)

ypread=mplr.predict(x_test)

print("Accuracy Score :",accuracy_score(y_test, ypread,normalize=True))
[01]
... Accuracy Score : 0.7213394380105934

```

Fig. 10. We were able to achieve an accuracy of 72.13% thanks to the utilization of "Logistic Regression."

B. Decision Tree (Classifier)

The "decision tree" comes from the realm of "supervised learning" and applies a non-parametric methodology. It serves both "classification and regression" purposes. "Decision trees" possess a root node, connecting branches, internal nodes, and leaf nodes, all assembled in a hierarchical manner. A "decision tree" commences with a root node, which isn't a branch of any other node. The internal nodes, often labeled as "decision nodes", are derived from the branches extending from the root node [19]. All nodes in decision trees employ evaluations using available attributes, aiming to create homogeneous subsets. These subsets are then represented by leaf nodes or terminal nodes. The leaf nodes of a "decision tree" signify all possible outcomes contained within the dataset. As an instance, you might use specific decision rules to decide whether to go surfing or not. The flowchart-style layout of such a decision-making model enables various teams within an organization to understand the rationale behind a decision [14]. The "divide and conquer" technique, using "greedy search" to determine optimal split points, underlies the learning process of a decision tree. Initially, the splitting technique divides records into different class labels. This top- down recursive method continues until the majority, if not all, of the records are sorted into homogeneous subsets. The complexity of the decision tree influences the likelihood of achieving pure leaf nodes, where each data point belongs to a single class. Smaller trees achieve pure leaf nodes more readily, indicating clearer categorization of data points. However, as a tree enlarges, maintaining this purity becomes challenging, often leading to data fragmentation and overfitting [21]. Hence, to reduce complexity and avoid overfitting, pruning is often employed, which involves trimming branches that split on insignificant attributes [8]. Cross-validation can then be used to evaluate the model's fit to the data. Several renowned decision tree algorithms, including ID3, C4.5, and CART, are built upon "Hunt's algorithm," conceived in the 1960s to emulate "human learning" within psychology. ID3, or "Iterative Dichotomiser 3," credits its invention to Ross Quinlan. This method employs entropy and information gain

as metrics to assess and identify potential splits in the "decision tree." C4.5, considered an improved version of Quinlan's earlier ID3, may use information gain or gain ratios as metrics to assess split points within decision trees. These measures enhance the accuracy and performance of the decision tree algorithm by identifying the most informative and efficient splits. CART, short for "classification and regression trees," was conceptualized by Leo Breiman. "Gini impurity" is typically used in this method to evaluate the optimal feature for splitting. It calculates the probability of misclassification for a randomly chosen attribute; lower values are preferable when utilizing "Gini impurity." It assists in identifying the feature resulting in the most accurate splits by examining the frequency of misclassification, thereby enhancing the algorithm's decision- making [18]. While there are various methods for choosing the best attribute at each node, "Gini impurity and information gain" emerge as the most commonly used splitting criteria in decision tree models. These metrics play an important role in determining the most informative attributes that drive the predictive power of the decision tree.

Entropy and Information Gain

To grasp the concept of information gain, it is important to delve into entropy. Information theory introduced the concept of entropy, which serves as a measure of the impurity or disorder within a set of sample values. By quantifying the uncertainty or randomness in the data, entropy provides valuable insights that contribute to understanding information gain. The following formula serves as its definition, where:

$$\text{Entropy}(S) = - \sum_{c \in C} (c) \log_2 (c)$$

- "S denotes the collection of data which helps determining the entropy".
- "c shows the classes that are present in the set of data".
- "The number of data points present divided by the total number of data points is represented by S p(c)".

In a dataset, entropy values typically range from 0 to 1, indicating various scenarios. A value of 0 denotes that there is only one class of samples in the dataset S. The entropy, on the other hand, achieves its maximum value of 1 when the dataset is evenly divided, with half of the samples belonging to one class and the other half to a separate class. The degree of entropy must be taken into account while choosing the optimum feature for splitting and building an ideal decision tree. The characteristic that produces the most informative splits is the one with the least entropy. The information gain, which indicates the difference in entropy before and after a split, calculates this [22]. Using the feature's target classification as a guide is the best method for categorising

the training data. We can accomplish the most efficient split and ultimately create the best decision tree model by choosing the best feature that maximises information gain. The following formula is typically used to determine “information gain”, where:

$$Information\ Gain(S, a) = Entropy(S) - \sum_{v \in values(a)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- “A display of a particular characteristic or class label”.
- “S |S_v|/ |S| denotes the ratio of the values in S_v to the total number of values in the dataset, and Entropy(S) is the entropy of the dataset”.
- “The entropy of the dataset, S_v, is entropy (S_v)”.

Despite being a straightforward algorithm, decision trees have the following benefits:

- **Interpretability**
We can see the model while using tree- based algorithms, which is one of their main advantages. You can see the algorithm's choices and how it categorized the various data pieces. This is an important advantage because most algorithms operate in a mysterious black box, making it difficult to understand why they predicted a particular outcome.
- **No pre-processing is required**
Numerous machine learning methods demand feature values to be as close as possible for the algorithm to understand feature changes' effects on the target as accurately as possible. Feature normalization, which places all features on the same scale and gives any change to those values the same proportionate weight, is the most typical pre-processing requirement. Instead of taking into account the full feature set, the rules in tree-based algorithms are constructed around each unique feature. Since just one feature is considered at a time, their values do not need to go through normalization.
- **Data robustness**
“Tree-based algorithms” make it simple to handle different data formats. They do not require pre-encoding of categorical variables because they can handle datasets that contain a mix of categorical and numerical data. The user's entire workflow is made simpler by these algorithms' natural capacity to handle such data preparation duties inside. Tree-based algorithms are useful for analysing and modelling datasets with a variety of data types because of their adaptability.

```

Decision Tree Classifier
└─>
    mpdc=make_pipeline(ct, dtc )
    mpdc.fit(x_train,y_train)

    ypread=mpdc.predict(x_test)

    print("Accuracy Score :",accuracy_score(y_test, ypread,normalize=True))
(418)
... Accuracy Score : 0.9382350380745129
    
```

Fig. 11 using a "decision tree classifier," we attained a 93.82% accuracy rate.

C. Random Forest Model

Leo Breiman and Adele Cutler created the "random forest machine learning method", which has become very well-known in the industry. By merging the outcomes of various “decision trees”, this algorithm applies an ensemble technique to get a unified result. Its user-friendliness and adaptability in managing “classification and regression” issues can be credited for its widespread use. Due to its capacity to harness the combined predictive potential of numerous “decision trees”, the “random forest method” has established itself as a useful tool, resulting in increased accuracy and robustness in a variety of applications.

In order to comprehend the "random forest model," it is helpful to first have a general understanding of the underlying "decision tree algorithm." An initial fundamental query, such as "Should I go surfing?" is posed using a decision tree. The response is then determined by asking a number of linked questions, such as "Is there a persistent swelling?" and "Is the wind strong at the beach?" The data is efficiently partitioned by these questions, which act as decision nodes inside the tree [23]. Each query influences the final judgement, which is symbolised by a leaf node. The "yes" branch is taken by observations that fit the requirements, whereas the alternative path is taken by observations that don't. The "Classification Regression Tree (CART) algorithm" is frequently used to train the decision tree technique, which seeks to evaluate the best splits within a subset of the data. The training process can use a variety of metrics, such as "Gini impurity," "information gain," or "mean squared error (MSE)." These measures make it easier to choose the splitting points that will maximise the decision tree's classification or regression accuracy.

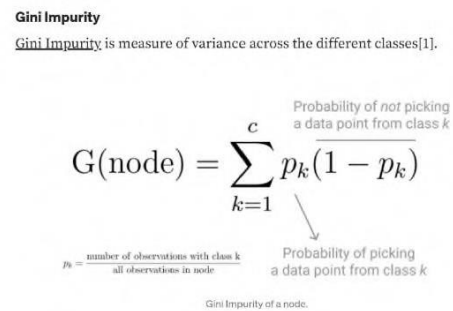


Fig. 12. Gini Impurity

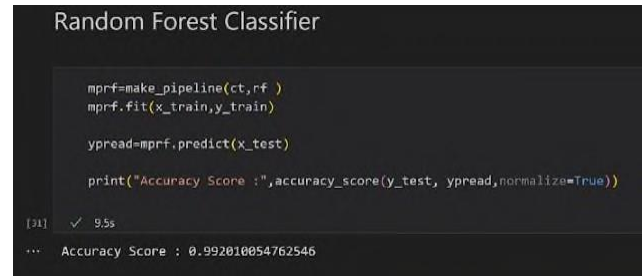


The “random forest model” aggregates the results from different “decision trees” to produce a consolidated prediction by using the “decision tree” method as the fundamental building block. The “accuracy and resilience of the model” are improved by using an ensemble technique. The popularity of the “random forest model” has grown due to its proficiency with a variety of datasets and efficient resolution of both classification and regression issues.

The "decision tree" belongs to the realm of "supervised learning" and applies a non-parametric methodology. It serves both "classification and regression" purposes. "Decision trees" possess a root node, connecting branches, internal nodes, and leaf nodes, all assembled in a hierarchical manner. A "decision tree" commences with a root node, which isn't a branch of any other node. The internal nodes, often labeled as “decision nodes”, are derived from the branches extending from the root node [19]. All nodes in decision trees employ evaluations using available attributes, aiming to create homogeneous subsets. These subsets are then represented by leaf nodes or terminal nodes. The leaf nodes of a "decision tree" signify all possible outcomes contained within the dataset. As an instance, you might use specific decision rules to decide whether to go surfing or not. The flowchart-style layout of such a decision-making model enables various teams within an organization to understand the rationale behind a decision [14]. The "divide and conquer" technique, using "greedy search" to determine optimal split points, underlies the learning process of a decision tree. Initially, the splitting technique divides records into different class labels. This top-down recursive method continues until the majority, if not all, of the records are sorted into homogeneous subsets. The “decision tree” complexities influences the likelihood of achieving pure leaf nodes, where each data point belongs to a single class. Smaller trees achieve pure leaf nodes more readily, indicating clearer categorization of data points. However, as a tree enlarges, maintaining this purity becomes challenging, often leading to data fragmentation and overfitting [21]. Hence, to reduce complexity and avoid overfitting, pruning is often employed, which involves trimming branches that split on insignificant attributes [8]. Cross-validation can then be used to determine the model's fit to the data. Several renowned decision tree algorithms, including ID3, C4.5, and CART, are built upon "Hunt's algorithm," conceived in the 1960s to emulate "human learning" within psychology. ID3, or "Iterative Dichotomiser 3," credits its invention to Ross Quinlan. This method employs entropy and information gain as metrics to assess and identify potential splits in the "decision tree." C4.5, considered an improved version of Quinlan's earlier ID3, may use information gain or gain ratios as metrics to assess split points within decision trees. These measures enhance the “accuracy and performance” of the “decision tree algorithm” by identifying the most informative and efficient splits. CART, short for "classification and regression trees," was conceptualized by Leo Breiman. "Gini impurity" is typically used in this method to determine the optimal feature for splitting. It calculates the probability of misclassification for a randomly chosen attribute; lower values are preferable when

utilizing “Gini impurity.” It assists in identifying the feature resulting in the most accurate splits by examining the frequency of misclassification, thereby enhancing the algorithm's decision making [18]. While there are

various methods for choosing the best attribute at each node, “Gini impurity and information gain” emerge as the most commonly used splitting criteria in decision tree models. These metrics play a role in determining the most informative attributes that drive the predictive power of the decision tree.



```

Random Forest Classifier

mprf=make_pipeline(ct,rf )
mprf.fit(x_train,y_train)

ypread=mprf.predict(x_test)

print("Accuracy Score :",accuracy_score(y_test, ypread,normalize=True))

[31] ✓ 9.5s
... Accuracy Score : 0.992816854762546

```

Fig. 13. With the help of a “random forest classifier,” we have been able to achieve 99.3% accuracy.

VIII. FINAL RESULT & ANALYSIS

The result analysis part of the study involves model selection and training. After data pre-processing and selecting the features, various ML algorithms such as “logistic regression”, “support vector machines”, decision trees”, and “neural networks” can be used to forecast the traces. The effectiveness of various algorithms can be evaluated using the dataset, and the best one can be identified based on maximized accuracy. The dataset needs to be separated into training and testing sets to evaluate the performance of the model. Techniques like “k-fold cross-validation” can be used to further evaluate the model. The model's adequacy of fit can be assessed using the Hosmer-Lemeshow test. The random forest classifier model has been preferred over other models, providing an accuracy of 99.3%. The model can be deployed in a real scenario by designing a web-based application or integrating it with an existing electronic health record system.

IX. CURRENT STATUS

Our model is currently working on a random forest classifier. We have tried using “logistic regression”, “decision tree classifier”, and “support vector machine” algorithms. However, the “random forest classifier” stands out from them all, and better scaling of data along with the algorithm provides us an accuracy of 99.3%. The accuracy result of other models that we have tried is also mentioned above. As we can see, the “logistic regression” model is providing 72.1% accuracy score, whereas the “decision tree classifier” model is providing an accuracy of 93.8%. Therefore, we have preferred the “random forest classifier model” over other ML models. Currently, we are only able to detect the traces of smoking inside our body with 0 or 1 as an output of the model. An output of ‘0’ represents the absence of traces of smoking while ‘1’ represents the presence. We are preparing a web-based platform for it to perform. The

screenshots provided below is showing the status of the website and the other

screenshot is showing the outcome page. Keenly observed, we have also provided the error messages column that is going to be guiding throughout the form fill-up process. Besides, the instruction column is also provided for guidance to a normal user.

Fig. 14. Screenshot of Result Page



X. EARLY WORKS

TABLE II

Early works

#	Author	Year	Previous Year Work	Algorithm used	Accuracy
1	Zen Zhan, Housing Chang, Xiao	2021	Research on smoking detection	Deep Learning	83%
2	Saurabh Sing, Pradeep Podder	2020	Real-Time Prediction for smoking activity	Machine Learning based multi-class classification model	74%
3	Mohammad Kharabshah, Omar Meqdadi	2019	Predicting Nicotine Dependence	Machine Learning	95.7%
4	Alessandro Ortis, Pasquale Caponnetto, Riccardo Polosa	2022	A Report on Smoking Detection and Quitting Technologies	Machine Learning	82.56%
5	Mona Lssabakhsh, Luz M	2022	Predicting Smoking Cessation among Us Adults	Random Forest, Generalised	5
6	Tzu-Chih Chien Chieh-Chuan Lin	2020	Smoking Behavior Detection	Deep Learning	6

XI. NOVELTY

The novelty of this work lies in the “detection of smoking traces in a human using the medical conditions” and providing a detailed view of their health condition. The project considers important parameters such as cholesterol, fasting blood sugar level, blood pressure, dental health, and a detailed view of the person’s lipid profile. The project aims to extend itself in the

future by detecting the concentration of nicotine or cotinine and determining heart disease and lung conditions. The “machine learning model” developed in this project has a 99% prediction performance and serves as an assistive tool for identifying patients at a higher risk of being smokers. The model can be deployed in a real scenario by designing a web-based application or integrating it with an existing electronic health record system. The “machine learning algorithm” used in this project is “random forest classifier”, which determines if an event’s probability of happening is depending on a given dataset of independent variables. The Hosmer-Lemeshow test evaluates the model’s adequacy of fit. The novelty of this work lies in the detection of smoking traces in a human using the medical conditions of the person and providing a detailed view of their health condition. The project considers important parameters such as cholesterol, fasting blood sugar level, blood pressure, dental health, and a detailed view of the person’s lipid profile. The project aims to extend itself in the future by detecting “the concentration of nicotine or cotinine” and determining “heart disease and lung conditions”.

XII. FUTURE POTENTIAL

We are not yet detecting the “concentration of nicotine or cotinine” present in our bodies. In the future, we can determine the “concentration of nicotine or cotinine”. We are also looking forward to determining the heart disease that might come up shortly according to the results that are collected from the individual tests. The condition of the lungs can be determined in future works on this project. The project seems to have a wide ground to extend itself for the betterment of livelihood and easier access to common people.

XIII. CONCLUSION

In this work, a “machine-learning model” was created to identify smoking residue in the body and rate the chance of developing heart disease. We collected a dataset with health, lifestyle, and demographic variables and pre-processed it by handling missing values, encoding categorical variables, and selecting important features. We used various machine learning algorithms, including “logistic regression”, “k- Nearest Neighbors”, “decision trees”, “artificial neural networks”, and “random forests”, to predict smoking traces based on health-related variables. Our model achieved a prediction performance of 99%, demonstrating its potential as a valuable tool for healthcare institutions to better understand and predict the likelihood of hospital admissions related to smoking. The novelty of our project lies in detecting smoking traces using medical conditions and providing a detailed view of the person's health condition, including “cholesterol, fasting blood sugar level, blood pressure, dental health, and lipid profile”. Our model can be deployed in a real scenario by designing a web-based application or integrating it with an existing electronic health record system.



XIV. ACKNOWLEDGEMENT

My deepest appreciation goes to everyone who contributed to this project, an invaluable learning journey. Profound thanks to Prof. (Dr.) Shantanu Sen and Chiranjib Dutta from Guru Nanak Institute of Technology for their support. Special gratitude to our mentors, Dr. Ananjan Maiti, and Mrs. Dola Saha, whose guidance and insightful feedback shaped our work. Kudos to the diligent Arijeet Roy and my supportive teammates who enriched this research with their efforts. Gratitude extends to our willing participants, without whom this endeavor wouldn't be possible. I honor the indispensable contributions of Chiranjib Dutta, Dr. Ananjan Maiti, and Mrs. Dola Saha once more

XV. REFERENCES

- [1] M. A. Serdar et al., "The correlation between the smoking status of family members and concentrations of toxic trace elements in the hair of children," in *Biological Trace Element Research*, vol. 148, no. 1, 2012, pp. 11-17.
- [2] E. Sazonov, P. Lopez-Meyer, and S. Tiffany, "A wearable sensor system for monitoring cigarette smoking," in *Journal of Studies on Alcohol and Drugs*, vol. 74, no. 6, 2013, pp. 956-964.
- [3] X. Zheng, J. Wang, L. Shanguan, Z. Zhou, and Y. Liu, "Smoky: Ubiquitous smoking detection with commercial WiFi infrastructures," in *Proc. 35th Annual IEEE International Conference on Computer Communications*, Apr. 2016.
- [4] S. K. Bashar and A. K. Mitra, "Effect of smoking on vitamin A, vitamin E, and other trace elements in patients with cardiovascular disease in Bangladesh: A cross-sectional study," in *Nutrition Journal*, vol. 3, no. 1, Oct. 2004.
- [5] U.S. Fire Administration, "Smoking-Related Fires in Residential Buildings," NFA [Online]. Available: <http://nfa.usfa.dhs.gov/downloads/pdf/statistics/v11i4.pdf>
- [6] Bedfont Scientific Ltd, "piCO+ Smokerlyzer," Bedfont Scientific Ltd [Online]. Available: <http://www.bedfont.com/cn/smokerlyzer/pico>
- [7] Y. Liu et al., "Detection of secondhand cigarette smoke via nicotine using conductive polymer films," in *Nicotine & Tobacco Research*, vol. 15, no. 9, 2013, pp. 1511-1518.
- [8] P. Wu et al., "Human smoking event detection using visual interaction clues," in *Proc. 20th International Conference on Pattern Recognition*, Aug. 2010.
- [9] A. A. Ali et al., "mPuff: Automated detection of cigarette smoking puffs from respiration measurements," in *Proc. ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN)*, Apr. 2012.
- [10] P. M. Scholl, N. Kücüküydiz, and K. V. Laerhoven, "When do you light a fire?," in *Proc. ACM conference on Pervasive and ubiquitous computing adjunct publication*, Sep. 2013.
- [11] "Effects of smoking on mortality," in *The Price of Smoking*, The MIT Press, 2004.
- [12] "6 Dixmier traces and positive traces," in *Theory*, De Gruyter, 2021, pp. 185-224.
- [13] "Multiple logistic regression," in *Applied Logistic Regression*, John Wiley & Sons, Inc., 2005, pp. 31-46.
- [14] A. Sagoolmuang and K. Sinapiromsaran, "Decision tree algorithm with class overlapping-balancing entropy for the class imbalanced problem," in *International Journal of Machine Learning and Computing*, vol. 10, no. 3, May 2020, pp. 444-451.
- [15] C. J. Mantas et al., "A comparison of random forest-based algorithms: Random credal random forest versus oblique random forest," in *Soft Computing*, vol. 23, no. 21, Nov. 2018, pp. 10739-10754.
- [16] "Use of a smartphone application on the detection of complications related to smoking," *Case Medical Research*, Aug. 2019.
- [17] S. S. Thakur, P. Poddar, and R. B. Roy, "Real-time prediction of smoking activity using machine learning based multi-class classification model," in *Multimedia Tools and Applications*, vol. 81, no. 10, Feb. 2022, pp. 14529-14551.
- [18] R. Tang, J. D. Robinson, and S. X. Day, "Personality traits predicting nicotine dependence and abstinence," *PsychEXTRA Dataset*, 2013.

- [19] "Kaggle: Your Home for DataScience." <https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/download?datasetVersionNumber=2> (accessed Jun. 09, 2023).
- [20] M. Issabakhsh et al., "Machine learning application for predicting smoking cessation among US adults," *Research Square Platform LLC*, Nov. 2022 [Online]. Available: <http://dx.doi.org/10.21203/rs.3.rs-2285331/v1>
- [21] "Smoking behavior," in *SpringerReference*, Berlin/Heidelberg: Springer-Verlag [Online]. Available: http://dx.doi.org/10.1007/springerreference_84592
- [22] T. P. Trappenberg, "Machine learning with sklearn," in *Fundamentals of Machine Learning*, Oxford University Press, 2019, pp. 38-65.
- [23] A. Ortis et al., "A report on smoking detection and quitting technologies," in *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, Apr. 2020.

XVI. BIOGRAPHIES



Arijeet Roy, an aspiring professional in the tech industry, completed his B.Sc. from the University of Calcutta and earned a "Master of Computer Applications (MCA)" from the "Guru Nanak Institute of Technology" in 2023.

His academic focus included research on mental health prediction and development, reflecting his unique interdisciplinary approach. Now embarking on his career, Arijeet has recently joined Capgemini as a software analyst, where his passion and theoretical expertise come into practical play.



Dr. Ananjan Maiti is a passionate young researcher who has a deep understanding of the subject of artificial intelligence based on disease detection. He is a passionate research content writer who always looks for new ways to express new ideas in reputed journals. He holds a Bachelor of Technology degree from JIS College of Engineering, with a specialization in Information Technology (2011).

He earned M.Tech from IIT Kharagpur, a prestigious institute with a specialization in "Information and Communication Technology (2015)". His scholastic odyssey did not end with the Master's degree; he encouraged His Excellency to pursue a "doctorate degree" in "Computer Science and Engineering" from the "University of Engineering & Management, Kolkata (2021)". He has a total of "8 years of teaching experience" and "2 years of industrial experience".

He has worked with the IEM Group, Techno Group, and JIS Group, all of which are reputable academic institutions in Kolkata. He is currently employed as an "Assistant Professor" in the "Department of Computer Science and Engineering" at "Guru Nanak Institute of Technology (JIS Group)". He has already published ten international papers in the previous three years, and his research interests include "medical image processing", "machine learning", "deep learning", and "the Internet of Things".



Prof. Dola Saha is a skilled academic with expertise in mathematics and computer science.

Her academic journey started at the Raja Peary Mohan College, University of Calcutta, where she earned a "Mathematics Honours degree" in 2004. She then achieved a "postgraduate degree" in "Computer application" from the "RCC Institute of Information Technology, MAKAUT", in 2008, followed by an M.Tech in "Information Technology" from "IEST, Shibpur", in 2016. Mrs Saha's areas of speciality include Cryptography, Steganography, Data Mining, and Machine

Learning, highlighting her broad and diverse skill set in information technology. As an active member of The Institution of Engineers (IE) and the Computer Society of India (CSI), she remains committed to her field, fostering professional relationships and making ongoing contributions.



Mr. Chiranjib Dutta is an accomplished academic and researcher with an impressive background in Mathematics and Computer Science. He embarked on his scholarly journey at the University of Calcutta, where he graduated with Honours in Mathematics. Broadening his knowledge base further, he pursued an MCA from Bangalore University, followed by an “ME in Computer Science and Engineering” from the “West Bengal University of Technology”.

Currently, Mr. Dutta is serving as the “Head of the Department of Computer Applications” at “Guru Nanak Institute of Technology, Kolkata”. He has dedicated his professional life to nurturing the next generation of tech-savvy individuals, imparting both knowledge and wisdom.

His primary research interest lies in the intriguing realms of Artificial Intelligence and Soft-computing. His insatiable curiosity and rigorous research methods have led to numerous publications in respected, peer-reviewed journals. These contributions not only highlight his in-depth understanding of the subject but also underline his unwavering commitment to advancing the field of computer science and its related application.

s