NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

# A Review on the Recent Advancements in Machine Learning-Assisted Tobacco Research

***Nabanita Ghosh* [*1], Krishnendu Sinha [2]**

[1] Assistant Professor in Zoology, Maulana Azad College, Kolkata-700013

[2] Assistant Professor in Zoology, Jhargram Raj College, Jhargram-721507

**Article Info**

***Keywords:*** *Machine learning; tobacco addiction; smoking abstinence; vaping; cancer agent; nicotine addiction*

*Received 22 February 2024*
*Revised 22 March 2024*
*Accepted 31 March 2024*
*Available online 20 May 2024*

https://doi.org/10.5281/zenodo.11223324



**Abstract**

*Tobacco smoking, a highly complex behavior influenced by genetic predisposition and environmental factors, is a grave global health alarm and a leading preventable cause of death, linked to severe diseases like osteoporosis and various cancers. Quitting rates remain dishearteningly low despite recommendations for nicotine replacement therapy and healthcare provider discussions. Presently, tobacco research generates vast data that can be harnessed by machine learning models, including supervised, unsupervised, and deep learning algorithms, which are gaining traction in tobacco research. This review delves into the intersection of traditional tobacco research and machine learning, elucidating the potential of ML methodologies in addressing the challenges of tobacco control. By leveraging diverse applications such as identifying smoking susceptibility predictors, predicting smoking behavior from genetic data, aiding cessation efforts with personalized predictors, automating data collection through apps, managing nicotine cravings, tracking smoking-induced diseases, monitoring illegal online vape sales, and assessing second-hand and third-hand smoke exposure, especially in infants, ML emerges as a powerful tool to combat the tobacco epidemic. Furthermore, this review highlights the significance of integrating ML techniques in tobacco research, emphasizing their role in advancing our understanding of smoking behavior, informing targeted interventions, and ultimately mitigating the public health burden associated with tobacco use. By harnessing the predictive power of ML algorithms, researchers and policymakers can tailor interventions to individual needs, optimize resource allocation, and accelerate progress toward a tobacco-free future.*

## 1. Introduction

Tobacco smoking is the primary preventable cause of death globally, resulting in the loss of 8 million lives annually, with 1.2 million of these being non-smokers exposed to secondhand smoke. [1–3]. Worldwide, this number will shoot up to 18.3 million by 2030 if a proper effort to limit tobacco smoking is not initiated timely [4–7]. According to a report by the World Health Organization (WHO), 36.7% of men and 7.8% of women, cumulatively 22.3% of the entire global human population smoked tobacco in 2020 [1]. Roughly an estimated 14% of all U.S. adults i.e., 34 million people are smokers, and they can expect to gain up to 10 years of life just by quitting

tobacco [8]. Hence prevention of smoking seems to be a critical step in controlling the tobacco smoking pandemic [6,9–12]. It has been shown that around 70% of smokers want to quit smoking while it takes an average of 6 quit attempts for them to quit before achieving persistent abstinence from smoking [8]. In such cases, a combination of nicotine replacement therapy (NRT) products and counseling has been proven to be effective [8,13].

Tobacco smoking is a rather complex learned behavior with a heritable genetic influence as high as 50% and maintained by physical nicotine dependence [6,8]. From socioeconomic status to education and from the environment to peer, numerous aspects have a significant influence on the development of tobacco smoking disorder [6,8,14]. However, in addition to traditional combustible smoking, electronic nicotine delivery systems (ENDS) are becoming a matter of concern for public health, especially in adolescents [15–22]. Though ENDS (like e-cigarettes and other 'vapes') was originally developed to aid individuals to quit smoking by allowing the controlled dose of nicotine, subsequent uncontrolled and unsupervised growth of the vaping market is imposing significant risk on the public health worldwide [15,17,22]. Vaping has emerged as a 'youth vaping epidemic' in the U.S. and intensified public health and regulatory inspection regarding this epidemic has directed more attention to uncontrolled advertising and trade of vaping goods on social media platforms [15,17,19–22].

A few years back computational methodologies were primarily centered on professional understanding, but nowadays, a transition from conventional methodology to machine learning (ML) models is evident and the credit goes to a large amount of data at disposal [6,23–25]. Currently, ML models are beginning to flourish in the fields of psychology, medicine, and public health [25]. Researchers are applying ML algorithms to address complex tobacco research-related questions like psycho-genetic predisposition on tobacco addiction/ tobacco smoking behavior, intricate behavioral patterns in smokers, deploying personalized digital assistance to quit smoking, tracking unregulated vaping trades over social media platforms, etc. [6,14,25]. In this review would like to appreciate these approaches. These were unimaginable a few years back but now coming into practice to control the tobacco epidemic, thanks to the next generation of ML algorithms and the great boom in big data research.

Furthermore, this review aims to provide a comprehensive examination of the current landscape of tobacco research and the evolving role of machine learning (ML) methodologies within this domain. By synthesizing recent literature and highlighting key findings, the review seeks to address several overarching research objectives. These include elucidating the significance of leveraging ML in tobacco research, identifying novel applications and advancements facilitated by ML algorithms, and assessing the implications of these developments on public health interventions and policy frameworks. By explicitly stating these research objectives, we aim to guide readers through the subsequent sections of the review, providing a clear framework for understanding the contributions and insights presented herein.

## 2. A Brief Account on Machine Learning

Machine Learning involves teaching computers to learn from data without explicit programming [26]. Unlike traditional programming, where specific rules are set, Machine Learning extracts patterns from data to create rules specific to each instance (Fig 1). Historically, it enables computers to learn without explicit programming [27]. Complex Machine Learning challenges fall into four core types: Classification, Regression, Clustering, and Rule extraction [28,29]. Classification labels discrete data points, Regression predicts numerical outcomes, Clustering groups similar data, and Rule extraction identifies associations among properties [29]. Machine

Learning systems are categorized as supervised, unsupervised, semi-supervised, or reinforcement learning based on human supervision during training [26,30,31].

### 2.1. Supervised learning

Supervised learning works with a group of labeled examples, i.e., training dataset. That can be mathematically represented as $\{(x_i, y_i)\}_{i=1}^{n}$, where an individual element $x_i$ is a feature vector and the dimensionality of the dataset is the length of this feature vector [30]. Each dimensional value is a feature, represented as $x_i^{(j)}$ [32]. Feature defines an instance. All the feature vectors are typically loaded into a matrix layout where an individual row signifies a vector for one instance and an individual column signifies all the instances' values for that feature. For an instance, $x_i$ in one collection characterizes an adolescent vape user, then the first feature $x_i^{(1)}$could involve his age, the second feature $x_i^{(2)}$, could involve his education, and so on. The label $y_i$ is generally an element to a finite set of classes or a real number indicating the label for the corresponding vector, e.g., ever-smoker or never-smoker [30,32]. A class can be conceptualized as a category to which a feature vector belongs. In supervised learning, the task of predicting a class is known as classification whereas, predicting a float is called regression [26,30]. Decision Trees (DTs) and Random Forests (RFs) (Fig 2), $k$-Nearest Neighbors ($K$NN), Support Vector Machines (SVMs) (Fig 3), Linear Regression (LR), artificial neural networks (ANN) (Fig 4) are some contextually relevant supervised learning algorithms. However, cumulatively it can be stated that a supervised learning algorithm intends to use a labelled dataset to generate a model that in succession takes an unlabeled feature vector $x_i$ as input and outputs a label for that [30].

### 2.2. Unsupervised learning

Unsupervised learning works with a set of unlabeled examples called a training dataset. Mathematically speaking, the training dataset is represented as $\{x_i\}_{i=1}^{n}$, where every element $x_i$ is a feature vector. Contrary to supervised learning, an unsupervised learning algorithm produces a model that accepts an unlabeled feature vector $x_i$ as input and either transforms it into another vector or into a scaler that can be employed to resolve a practical problem [26,30]. Clustering is a critical unsupervised learning technique appropriate for locating groups of analogous objects in a large pool of objects. A clustering model sends the ID of the cluster for every feature vector in the dataset. In other words, it puts a label/ID on unlabeled data points based on its features [30]. Hierarchical Cluster Analysis (HCA), $K$-Means, Isolation Forest, t-Distributed Stochastic Neighbor Embedding (t-SNE), and Principal Component Analysis (PCA) are a few vital unsupervised learning clustering algorithms [26]. Instances of unsupervised learning in tobacco research include discovering the topics of conversations related to tobacco on social media or finding possible nicotine addiction subtypes by evaluating the MRI data of brain addicts [33].

## 3. ML in Prognosis of Tobacco Smoking Induced Ailments

### 3.1. Hierarchical clustering and SVM in Smoking-induced lung cancer prognosis

Lung cancer ranks among the leading causes of cancer-related deaths globally due to its late detection, high malignancy rate, and rapid progression [34,35]. The duration of tobacco smoke exposure, rather than pack size, strongly correlates with lung cancer risk, with active smoking years significantly increasing the odds of developing the disease [35–37]. Despite most cases being linked to tobacco smoke, approximately 10-15% of U.S. lung cancer cases occur in never-smokers, suggesting other contributing factors beyond smoking [34,35]. Chapman et al.'s meta-analysis

highlighted key driver mutations and patient characteristics such as sex, ethnicity, and smoking status as common in never-smoker lung cancers [35]. Recent work by Chen and Lin identified five feature risk pathways capable of effectively distinguishing smokers with lung cancer from those without, facilitating tailored treatments, and identifying high-risk individuals [34]. Their study analyzed standardized expression profiles from the GEO database, identifying numerous differentially expressed genes (DEGs) using the limma algorithm [34]. Hierarchical clustering confirmed DEGs' effectiveness in classifying smoker subgroups and revealing similarities [34]. GO and KEGG enrichment analyses mapped DEGs to a protein-protein interaction network, highlighting functions associated with up-and down-regulated genes. Five optimal feature pathway sets were identified, improving diagnostic accuracy to 84% using an SVM model, which further increased to 90% with clinical data integration [34]. The study concluded that smokers with prolonged smoking history, lymphadenopathy, and larger nodules are at higher risk of developing lung cancer.

While recent studies demonstrate the potential of ML in lung cancer risk assessment, it's important to acknowledge limitations. For instance, relying solely on standardized expression profiles from the GEO database may introduce biases. Additionally, while ML models show impressive diagnostic accuracy, concerns about overfitting and generalizability to diverse patient populations remain. Furthermore, ML algorithms often lack interpretability, hindering understanding of underlying biological mechanisms. Future research should focus on validating models with diverse datasets, integrating clinical and molecular data, and enhancing interpretability for effective clinical translation.

### 3.2. XGBoost and Smoking-induced non-communicable diseases prognosis

Noncommunicable diseases (NCDs) are posing a significant death threat to the global population with a staggeringly high rate of 70% of all global mortalities [38]. Tobacco has emerged as a key etiological agent for NCDs. Smoking is also jeopardizing the United Nations' Sustainable Development Goals (SDGs) while cessation of smoking could help to reduce the global NCD burden to one-third by the end of 2030 [39]. Smoking-induced noncommunicable diseases (SiNCDs) include diabetes, stroke, all heart disease, cancers, chronic respiratory diseases, Parkinson's disease, Crohn's disease, etc. Death rates due to SiNCDs could be effectively checked by early diagnosis, efficient treatment, and cessation of smoking [34,38,40,41].

ML has shown great prospects in forecasting, assessing, and early diagnosis of SiNCDs [38]. In a recent study, Davagdorj et al. developed a framework based on XGBoost to predict SiNCDs [38]. The study used the National Health and Nutrition Examination Survey datasets of South Korea (KNHANES) and the United States (NHANES) to build the optimal XGBoost model for SiNCDs predictions [38]. They also compared the proposed model with the random forest, logistic regression, SVM, support vector machine recursive feature elimination (SVM-RFE), random forest-based feature selection (RFFS), *K*NN, MLP, and neural network baseline classifiers [38]. The XGBoost-based framework has been combined with the hybrid feature selection (HFS) method for SiNCD prediction and, selected features are fed into the model [38]. Primarily, HFS is accomplished in three stages picking important features by t-test and chi-square test, obtaining dissimilar features by multicollinearity analysis, and finally selecting the best representative features employing the least absolute shrinkage and selection operator (LASSO) [38]. The study revealed the most characteristic features of SiNCDs in the general populations of the United States and South Korea [38]. Some important predictors that emerged from this study are obesity, overweight, alcohol intake, psychological health, blood cholesterol level, age, number of smokers living in the same house, etc. [38]. The suggested model delivered critical features to expand the interpretability of the SiNCDs prediction model and is likely to aid in the timely diagnosis and prevention of SiNCDs in public health issues [38].

While Davagdorj et al.'s study on XGBoost-based SiNCD prediction models shows promise, it's important to critically assess its strengths and limitations. While achieving high predictive accuracy, reliance on observational datasets like KNHANES and NHANES may limit generalizability. Comparative analysis with baseline classifiers offers insights, yet further investigation into feature selection methods' impact on interpretability and generalizability is needed. Key predictors identified underscore SiNCDs' multifactorial nature, prompting the integration of additional data sources for improved accuracy and clinical utility. By critically evaluating ML approaches, researchers can advance effective public health interventions targeting SiNCDs.

### 3.3. SVM-RFE and RF in Smoking-related postmenopausal osteoporosis management

Osteoporosis is a skeletal disorder identified by diminished bone strength and bone mineral density, bone fragility, altered bone microstructure, etc. [42,43]. Due to changed hormonal status, post-menopausal women are especially vulnerable to osteoporosis, as more than 50% of them develop a condition called post-menopausal osteoporosis (PMOP) [44]. B-lymphocytes severely influence PMOP by generating cytokines that control the activity of osteoblast and osteoclast and downregulating own MAPK3 and ESR1 which alter the secretion of factors leading to increased osteoclastogenesis and decreased osteoblastogenesis [42,45]. However, smoking has emerged as an independent risk factor for PMOP which helps to develop the disease with twice the frequency in female smokers to non-smokers, leading to a disorder termed, smoking-related postmenopausal osteoporosis (SRPO) [45–47].

A current study has recognized six genes (HNRNPC, TCEB2, PFDN2, RPS16, PSMC5, and UBE2V2) as prospective biomarkers for SRPO by analyzing gene expression patterns of 20 postmenopausal female smokers' circulating B lymphocytes with low or high BMD employing, weighted gene co-expression network analysis (WGCNA), random forest (RF), support vector machine-recursive feature elimination (SVM-RFE), protein-protein interaction (PPI) and functional analyses, as well as ROC curve analysis [42]. The study used the GSE13850 microarray dataset from Gene Expression Omnibus (GEO) to identify gene modules related to SRPO by applying protein-protein interaction (PPI), analysis weighted gene co-expression network analysis (WGCNA), and pathway functional enrichment analyses [42]. SVM-RFE and RF selected feature genes [42]. The study detected eight highly conserved modules in the WGCNA network and the strongly SRPO correlated genes in the module that was used for building the PPI network, where a total of 113 hub genes in the core network has been identified using topological network analysis closely related to ATPase activity, regulation of RNA transcription and translation and immune-related signaling [42]. The study recognized prospective genetic biomarkers and offered a new understanding of the fundamental molecular mechanism of SRPO which might offer a narrative intuition into the diagnosis and treatment of SRPO [42].

The study identifies six genes as potential biomarkers for smoking-related postmenopausal osteoporosis (SRPO) using advanced techniques like weighted gene co-expression network analysis (WGCNA) and protein-protein interaction (PPI) analysis. While offering valuable insights into SRPO pathogenesis, limitations include reliance on a single dataset and the need for further validation. Nonetheless, these findings pave the way for future research aimed at improving diagnostic and therapeutic strategies for SRPO.

### 3.4. Ensemble learning in Smoking and HIV prognosis

Despite over 60% of HIV-infected individuals being smokers, smoking remains an overlooked independent risk factor for adverse outcomes in this population [48–50]. Epigenome-wide association studies (EWAS) have identified multiple DNA methylation CpG sites in white blood cells (WBCs) linked to smoking-related traits and diseases, including cancer, heart diseases, and

chronic inflammation [51–56]. While these findings primarily stem from HIV-uninfected populations, it's known that HIV infection affects the host epigenome and serves as a predictor for HIV-related aging and cognitive impairment [57–60]. Zhang et al. utilized ensemble machine learning models to uncover 698 DNA methylation sites associated with smoking in HIV-infected WBCs, enabling the prediction of frailty and mortality in this population [58]. They demonstrated the efficacy of DNA methylation-based machine learning in HIV prognosis. The study identified 137 epigenome-wide significant CpGs for smoking in WBCs from 1137 HIV-positive individuals, of which 698 were selected as predictors of high frailty, showing robust performance in both training and independent samples. Additionally, the study revealed a correlation between a DNA methylation index constructed from these CpGs and a 5-year survival rate [HR = 1.46; 95%CI 1.06~2.02, p = 0.02] [58]. Pathway analysis highlighted significant enhancement in canonical integrin signaling pathways and other non-canonical pathways related to organ injury, cancer, and abnormalities, suggesting the biological relevance of features chosen by ensemble learning in smoking-related diseases. These findings were based on DNA samples extracted from HIV-infected individuals' WBCs in the Veteran Aging Cohort Study (VACS).

Zhang et al.'s study using ensemble machine learning to identify DNA methylation sites associated with smoking in HIV-infected white blood cells represents a significant advancement in understanding smoking's impact on HIV prognosis. While the study demonstrates the efficacy of DNA methylation-based machine learning in predicting frailty and mortality in this population, it is crucial to acknowledge limitations such as reliance on WBC DNA samples and potential confounding factors. Further validation in independent cohorts is needed to confirm the findings' generalizability. Nevertheless, these results highlight the potential of machine learning in uncovering novel biomarkers and pathways relevant to smoking-related outcomes in HIV-infected individuals, offering insights for personalized interventions and clinical management strategies.

## 4. ML and Indirect Tobacco Smoke Exposure

### 4.1. Assessing fetal exposure to maternal smoking by supervised ML

Smoking during pregnancy profoundly impacts infant health and is associated with adverse outcomes [61]. Despite this knowledge, only 45% of mothers who smoked before pregnancy ceased smoking during gestation [62]. In utero exposure to maternal smoking increases the risk of non-communicable diseases (NCDs) such as pediatric obesity and nicotine addiction in offspring, primarily through epigenetic modifications like DNA methylation [62–67]. These alterations can persist from infancy to adulthood and possibly beyond [68–70]. Markunas et al. conducted an epigenome-wide association study (EWAS) on 889 infants, identifying DNA methylation changes associated with maternal smoking during the first trimester [62]. They used robust linear regression models in R to analyze CpG sites related to maternal smoking status, identifying 185 CpGs with significant methylation changes linked to 110 gene regions [62]. Notably, they identified 10 previously unknown genes associated with smoking exposure, including ATP9A and FRMD4A, which are implicated in smoking cessation and embryonic development [62,71]. Their findings suggest potential age-dependent differences in methylation patterns and raise questions about the long-term health effects of maternal smoking exposure [62]. In another study, a DNA methylation score for in utero tobacco smoke exposure was developed using supervised machine learning algorithms [66]. This score, validated in multiple cohorts, can help identify individuals exposed to maternal smoking during pregnancy based on DNA methylation data. It provides a valuable tool for studies lacking direct maternal smoking data, aiding in the adjustment of models to account for its effects.

In the section discussing maternal smoking during pregnancy, machine learning (ML) techniques aid in identifying DNA methylation changes associated with smoking exposure. While ML offers

promise in uncovering molecular mechanisms and developing predictive models, it's important to acknowledge limitations, such as potential confounding factors. Further research is needed to elucidate the functional significance of these changes and their long-term health implications for offspring.

### 4.2. Assessing secondhand and thirdhand tobacco smoke exposure to infants by multivariable linear regression models

As discussed throughout this review, a considerable amount of work has been done on firsthand tobacco smoke exposure, there is less appreciation of secondhand tobacco smoke exposure and ever lesser consideration for thirdhand tobacco smoke exposure [72–76]. Exposure to the smoke discharges from the burning end of cigarettes and smoke respired by smokers are considered secondhand smoke exposure whereas, exposure to the tobacco-related gases and particles that become fixed in materials like furniture, carpets, toys, cloths, beds, upholstery, etc. are considered as thirdhand tobacco smoke exposure [75,76]. Infants are mostly affected by these kinds of involuntary secondary exposures. Parks et al. assessed secondhand and thirdhand tobacco smoke exposure in infants residing in Canada [72]. The team tried to recognize the sources of tobacco smoke exposure using ML methods combined with prediction modeling for the infants from the CHILD Cohort study, a four-center (Manitoba, Edmonton, Toronto, and Vancouver) longitudinal population-based birth-cohort study, that registered 3455 mother-child duos in-between 2008 and 2012 [72]. Urinary nicotine biomarkers, cotinine, and trans-3'-hydroxycotinine (3HC) were used to assess tobacco smoke exposure in infants [72]. The study evaluates the capability of questionnaires to predict the variability of urinary cotinine and 3HC concentrations among 2017 3-month-old infants by using multivariable linear regression models [72]. These models were selected with the help of conceptual and data-driven strategies involving random forest regression. The team detected cotinine and 3HC in the urine samples of 76% and 89% of the infants, respectively, though, only 2% of mothers testified smoking before and during their pregnancy [72]. Questionnaire-based models explained 41% and 31% of the variance in 3HC and cotinine levels, respectively [72]. Detected concentrations recommend 0.25 and 0.50 ng/mL as cut points in cotinine and 3HC to characterize secondhand smoke exposure and this proposes that 23.5% of infants had modest or consistent smoke exposure [72]. This suggests that parents' efforts to decrease tobacco smoke exposure to their infants are being compromised due to a lack of consideration of the universality and persistence of secondhand and thirdhand smoke. Parks et al. could not predict over half of the variation in urinary 3HC and cotinine in infants through their model probably because of the ubiquity of thirdhand smoke and the possibility for oral and dermal routes of nicotine exposure [72]. These predictors need to be evaluated and assessed in future studies for a better-fit model.

While the review provides insights into the assessment of secondhand and thirdhand tobacco smoke exposure in infants using machine learning (ML) methods, a deeper examination of the strengths, limitations, and implications of these approaches is warranted. The study by Parks et al. sheds light on the challenges of accurately predicting tobacco smoke exposure in infants, despite the utilization of ML techniques combined with prediction modeling. The observed discrepancy between questionnaire-based predictions and urinary biomarker levels underscores the complexity of assessing tobacco smoke exposure in real-world settings. Moreover, the inability to predict a significant portion of the variance in urinary biomarker levels highlights the limitations of current ML models in capturing the full spectrum of exposure pathways and sources, particularly those associated with thirdhand smoke. These findings emphasize the need for further research to refine ML models and evaluate additional predictors that account for the ubiquity and persistence of thirdhand smoke exposure. Additionally, future studies should explore the implications of oral and dermal routes of nicotine exposure in infants, enhancing the development of more accurate and comprehensive models for assessing tobacco smoke exposure in this vulnerable population.

## 5. ML and Unconventional Sources of Nicotine

Nicotine in any form is dangerous to both physical and mental health. Unfortunately, besides conventional sources of combustible nicotine products like cigarettes, cigars, pipes, etc., unconventional sources of nicotine cumulatively known as electronic nicotine delivery systems (ENDS), are getting immensely popular, especially among teenagers [77–80]. ENDS are popularly called vapes which include a wide variety of products like e-cigarettes, e-cigs, vaporizers, vape pens, e-pods, e-pipes, etc., and are prevalently marketed in the U.S. by brands like JUUL, Puff Bar, SMOK, and Vuse [77–79,81,82]. E-cigarettes were launched into the US market in 2007 and have swiftly become a common source of nicotine for several patients [82]. These are constructed to imitate smoking by heating a nicotine-containing solution, called 'e-liquid', producing an aerosol that the user inspires. Though the short- and long-term effects of an e-cigarette are still controversial with certain negative acute effects on blood pressure, heart rate, and airway resistance, their use is growing alarmingly [80,82]. Though vapes were originally designed to aid long-term tobacco smoking cessation by delivering a regulated amount of nicotine, their marketing, and use took a wrong turn by popularizing among teens and youth as a style statement. The use of vapes in nicotine withdrawal therapy is not recommended by FDA [77,79]. Vapes are also not proven to do any good in assisting nicotine cessation but it is affecting a large amount of youth worldwide. The annual National Youth Tobacco Survey in 2021 found that more than 2 million youth are currently using e-cigarettes in the U.S. of which 8% of middle school students and 28% of high school students are involved [77,78,81–85]. Among them, 40% use e-cigarettes frequently, and 25% use e-cigarettes daily [79,83]. Nearly 85% of teens choose flavored vapes [83]. These disturbing figures are indicative of a strong vape-induced nicotine dependency [83]. Using nicotine as a teen may have a lasting effect on attention, memory, and learning that supports future addiction to nicotine [77,78,80,83,84]. More youths who start taking nicotine via vaping ultimately end up in combustible cigarettes [77,79,83,84]. Along with nicotine vapes contain toxic metals like nickel, chromium, and lead which are proven lung-damaging substances [86–88]. Several of the chemicals found in combustible cigarette smoke like acetaldehyde, formaldehyde, and acrolein, are also found in many e-cigarette aerosols, and breathing in these chemicals causes irreversible lung injury [77–79,83–85].

Vaping-associated pulmonary injury (VAPI) or e-cigarette or vaping product use-associated lung injury (EVALI) was an outbreak in 2019 in the U.S. which is considered an acute or subacute respiratory illness characterized by a range of clinicopathologic findings imitating several pulmonary diseases [89]. The epidemic set out in March 2019 when a collection of incidents appeared in the USA of patients who had developed lung injuries related to the use of vapes [89,90]. As of February 2020, more than 2800 patients have been affected by EVALI and been admitted to various US hospitals of which 68 deaths have been reported so far [89,90]. Beyond the US, Canada and Europe have reported a few cases of EVALI with fatal outcomes [91,92]. Bestowing to the CDC criteria, clinical diagnosis of EVALI entails the usage of an e-cigarette in the 90 days before the onset of early symptoms, pulmonary infiltrates on basic chest CT or radiograph, and lack of any additional possible etiology, such as infection, including acute fibrinous pneumonitis, diffuse alveolar damage, or organizing pneumonia, typically bronchiolocentric and complemented by bronchiolitis [89–92]. Reports also find the presence of vitamin E acetate, an additive of many vape liquids, in the bronchoalveolar lavage (BAL) of most EVALI patients which makes health officials believe is the primary, but not the only, cause of EVALI [90].

Hence, alongside controlling conventional sources of tobacco smoking, controlling these unconventional sources poses challenges to regulating authorities like the FDA and in such scenarios, ML algorithms are proven indispensable.

### 5.1. Assessing ENDS uses among adolescents by supervised ML

Intervention strategies to stop teenagers from using ENDS should be based on robust predictors of ENDS use which typically vary from predictors of conservative combustible tobacco product use [32]. Studies find novel emerging predictors like orientation towards digital media use, orientation towards new technologies, etc., in adolescents which is different from traditional tobacco smoking predictors [93,94]. As discussed earlier, identifying the novel and emerging predictors for ENDS use will help healthcare providers and law enforcement agencies combat this ENDS epidemic. Han et al. identified emerging predictors for adolescent ENDS uses employing supervised ML algorithms, specifically the penalized logistic regression algorithm due to its analytical benefits such as high prediction performance and decent model interpretability in contrast to other ML algorithms [32,95]. Though they also checked three other well-established algorithms, namely distributed random forest, gradient boosting machine, and deep neural networks, none were found superior to the penalized logistic regression algorithm [32]. The team investigated nationally representative multi-wave longitudinal survey data (2013–2018) taken from the Population Assessment of Tobacco and Health Study (PATH) [32]. A sample of juveniles and teenagers (12–17 years) who completed Wave 2 (n= 7958), Wave 3 (n = 6260), and Wave 4 (n = 4544) and never used any tobacco products at baseline were studied [32]. The penalized logistic regression evaluates self-reported past-month ENDS use (i.e., current use) at Waves 2–4 depending upon the variables measured at the former wave [32]. The algorithm demonstrated a proper ability to distinguish between wave-wise ENDS uses and non-uses depending on the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) [32]. From the study, frequent social-media usage appeared as a critical variable in guessing adolescent ENDS [32]. Such social media use variables are seldom significant predictors for other substance use, specifically for adolescent substance use behaviors, and are completely ineffective as conventional smoking behavior predictors. Authors rightfully predicted that digital media use emerging as a prominent predictor of ENDS uses for adolescents because increased ‘technophilia’ is exposing youth to lucrative and aggressive campaigns of ENDS over the internet and social media [32,93]. Alarmingly it has been recovered from the study that social media use variables were also capable of forecasting cigarette smoking in subsequent years (i.e., Wave 4 outcome) as long-term usage of ENDS might make the youth susceptible to combustible tobacco products [32,93,94]. The study also pointed out that, adolescents who developed a curiosity about ENDS across social media might try out alcohol or marijuana along with ENDS, which makes them susceptible to these substances' uses [32,93].

Furthermore, while discussing the study by Han et al. on the identification of predictors for adolescent ENDS (electronic nicotine delivery system) use, it is crucial to acknowledge the methodological strengths and limitations of employing supervised machine learning algorithms. The use of penalized logistic regression demonstrated high prediction performance and model interpretability, which are commendable attributes for informing intervention strategies. However, it's essential to recognize that the choice of algorithm may influence the results obtained, and alternative algorithms, such as distributed random forest and gradient boosting machine, should also be considered for comparison to ensure robustness. Additionally, the reliance on self-reported data for ENDS use may introduce biases, such as underreporting or social desirability bias, which could impact the accuracy of the predictive models. Moreover, while the study identifies social media usage as a significant predictor for adolescent ENDS use, further exploration is warranted to understand the underlying mechanisms driving this association and its implications for targeted interventions. Despite these limitations, the findings underscore the importance of considering novel predictors, such as digital media use, in designing intervention strategies to address the emerging ENDS epidemic among adolescents.

### 5.2. Vaping-related surveillance on Twitter and Instagram by stacking ensemble learning and supervised ML

Social media has become a popular medium to express and share thoughts, experiences, and opinions about almost every issue an individual can think of. Twitter is a highly accepted social media platform with an enormous user base and hence has effectively been used for health-related public surveillance in the areas of mental health and well-being, illegal drug use, and other health-related issues [96–99]. Public opinion varies widely about vaping and those opinions often get reflected on Twitter. Twitter encompasses rich data communicated by users about their behaviors and experiences, as well as thoughts on vaping [99]. Hence Twitter can also be used as an excellent source for vaping-related data mining. These data can be effectively used to detect users at risk of vaping-related adverse health conditions and thus benefit from an intervention regarding vaping cessation. In a recent study, Ren et al. developed a stacking ensemble learning model to routinely capture vaping-associated tweets and their accompanying user accounts by assessing millions of tweets [14]. An enormous number of tweets was the main obstacle to such surveillance. This work could provide an excellent tool for detecting and intervening in probably vulnerable individuals concerning ENDS usage. The team has applied seven well-established ML and deep learning algorithms namely Naïve Bayes, Random Forest, XGBoost, stacking and voting ensemble, models Support Vector Machine, Multilayer Perception, and Transformer Neural Network, and for their custom-built classification model [99]. They mined a set of sample tweets in the 2019 EVALI outbreak using the Twint Python package and generated an annotated data set to train and evaluate these models [99–101]. The stacking ensemble learning accomplished the peak performance with an F1 score of 0.97 and was selected as the optimum model for the study [99]. The study provides a good ensemble classifier with a stacking method to be used for screening and mining millions of tweets to detect persons, speaking and networking regarding vaping on social media sites and to extend to individuals who might be at risk for serious health consequences owing to vaping and could be benefited from unswerving link to quit support and associated intervention programs [99]. In another recent study, by using Targeted Topic Modelling (TTM) (a supervised machine learning algorithm), Costigina et al. found commercial JUUL-related Instagram posts highlighting e-cigarette-related marketing [102]. These posts have convincing messages promoting vape trial and use, leading to nicotine dependence, especially among vulnerable adolescents.

Moreover, while the utilization of social media data and machine learning (ML) techniques holds promise for surveillance and intervention efforts in tobacco control, it is essential to critically examine the strengths, limitations, and implications of these approaches. Despite the demonstrated effectiveness of ML algorithms such as stacking ensemble learning and targeted topic modeling in identifying vaping-related content on platforms like Twitter and Instagram, several challenges persist. For instance, the reliance on publicly available social media data raises concerns about representativeness and biases inherent in the data, potentially limiting the generalizability of findings. Additionally, the rapid evolution of social media platforms and user behaviors necessitates ongoing refinement and adaptation of ML models to ensure their continued efficacy in detecting emerging trends and patterns related to tobacco use. Furthermore, ethical considerations surrounding user privacy and consent must be carefully addressed in the collection and analysis of social media data for public health purposes. By critically evaluating the strengths and limitations of ML approaches in leveraging social media data for tobacco research, researchers can better navigate the complexities of this burgeoning field and maximize the utility of these innovative methodologies in advancing tobacco control efforts.

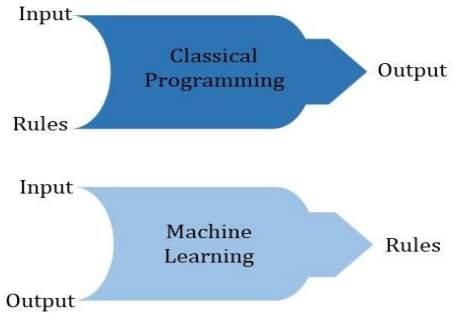


**Fig 1.** The fundamental idea behind classical programming and machine learning methods.

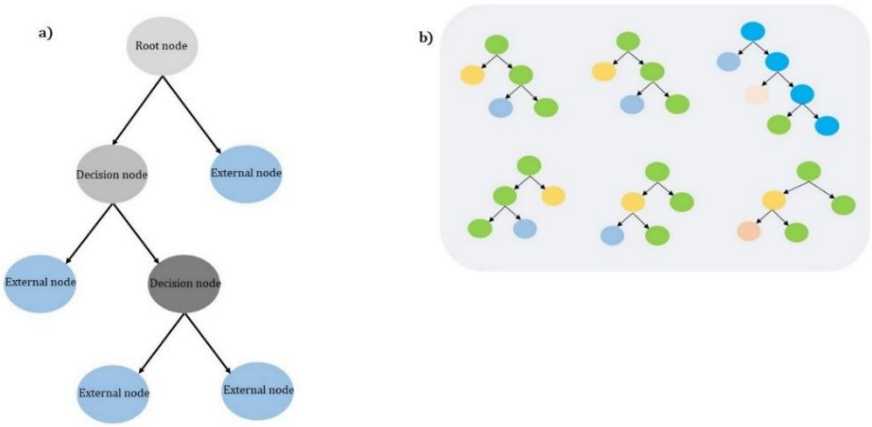


**Fig 2.** Decision Trees, (a) fundamental structural components of a decision tree; (b) an Ensemble of six distinct decision trees.

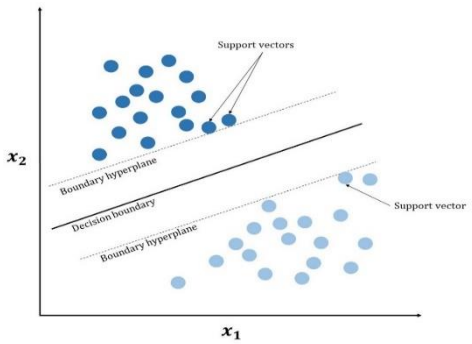


**Fig 3.** SVM is represented graphically showing hyperplanes and support vectors. The ideal hyperplane splits the data points into two categories characterized by grey and blue dots.

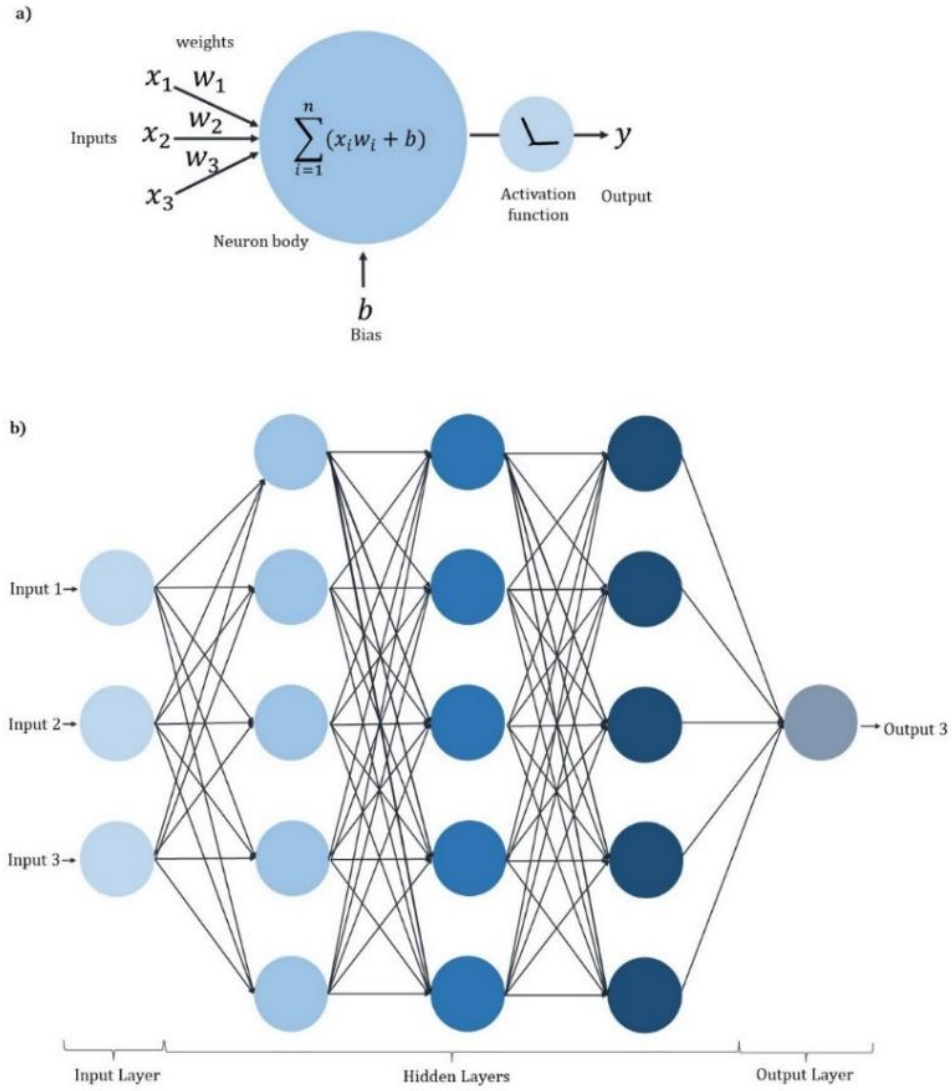


**Fig 4.** ANN, (a) single rectified linear unit (i.e., single neuron plus rectified function as its activation function) with matching weights ($w$) and bias ($b$); (b) 3-layer deep stack of dense layers with input-output layers.

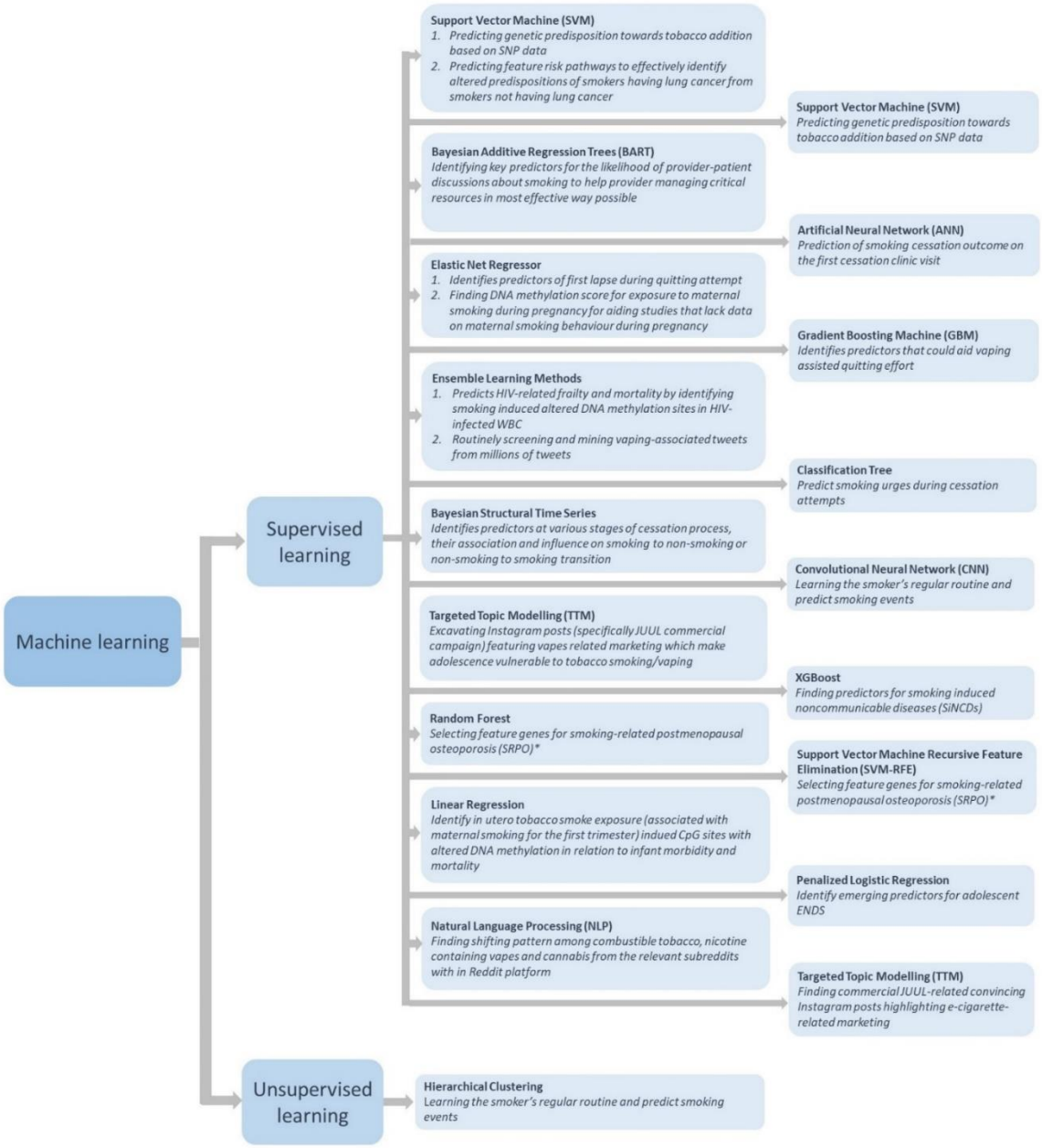


**Fig 5.** An overview of the divisions of ML and related algorithms and their utilization in solving tobacco research-related issues.

## 6. Conclusion

The effective and ethical integration of ML holds the potential to revolutionize tobacco control interventions, especially in light of the burgeoning volume of available data. ML is transforming various aspects of tobacco research, including the analysis of epigenetic and genetic data to predict susceptibility to tobacco addiction and related illnesses (Fig 5). Additionally, ML aids in the extraction of pertinent smoking-related data from social media platforms such as Twitter, Instagram,

and Reddit [103,104]. ML facilitates comprehensive scrutiny of tobacco use by automatically analyzing social media content, offering insights into public opinions on smoking, including contentious topics like vaping. This data enables the identification of factors contributing to adolescent tobacco use, thereby informing targeted intervention strategies for high-risk groups. Moreover, ML assists in evaluating secondhand and thirdhand tobacco exposure in infants, aiding in the identification and mitigation of exposure sources [104]. From internet surveillance to predictive analytics based on various data sources, ML plays a pivotal role in advancing tobacco research. However, it is crucial to acknowledge the opacity of many ML models, requiring careful consideration of the statistical methods and computational tools used to interpret findings accurately. Such findings often capture complex interactions among predictors and outcomes that conventional statistical methods may struggle to uncover [104]. Tobacco researchers must also address inherent biases in ML applications to tobacco control and work towards mitigating them effectively. By staying abreast of emerging trends and responsibly harnessing sophisticated tools, ML has the potential to unlock new frontiers in tobacco research.

Furthermore, synthesizing findings from diverse studies allows for the identification of overarching themes and patterns essential for advancing tobacco research and intervention strategies. While the discussion underscores the transformative potential of ML in tobacco control, a more critical analysis of the evidence presented can enrich our understanding of its implications. By rigorously evaluating the strengths and limitations of ML applications in tobacco research, researchers can pinpoint areas for refinement and future exploration. Moreover, delving into the practical implications of these findings can inform evidence-based decision-making in tobacco control efforts. Specifically, highlighting gaps in the literature and proposing avenues for further investigation can guide future research endeavors. Future studies might focus on refining ML algorithms to address inherent biases, enhancing data quality and interpretability, and exploring innovative applications in tobacco prevention and cessation programs. Additionally, fostering interdisciplinary collaborations between researchers, policymakers, and public health practitioners is paramount for translating ML-driven insights into actionable public health interventions. Embracing these recommendations, the tobacco research community can fully leverage ML's potential to effectively combat the tobacco epidemic.

## 7. Funding details

This research received no funding.

## 8. Conflicts of interest

The authors report there are no competing interests to declare.

## References

[1] Tobacco, https://www.who.int/news-room/fact-sheets/detail/tobacco.

[2] M. B. Reitsma et al., Spatial, Temporal, and Demographic Patterns in Prevalence of Smoking Tobacco Use and Attributable Disease Burden in 204 Countries and Territories, 1990–2019: A Systematic Analysis from the Global Burden of Disease Study 2019, The Lancet 397, 2337 (2021).

[3] X. Dai, E. Gakidou, and A. D. Lopez, Evolution of the Global Smoking Epidemic over the Past Half Century: Strengthening the Evidence Base for Policy Action, Tob Control 31, 129 (2022).

[4] C. W. Warren, N. R. Jones, M. P. Eriksen, and S. Asma, Patterns of Global Tobacco Use in Young People and Implications for Future Chronic Disease Burden in Adults, Lancet 367, 749 (2006).

[5] Current Cigarette Smoking Among Adults — United States, 2011, https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6144a2.htm.

[6] Y. Xu et al., Prediction of Smoking Behavior From Single Nucleotide Polymorphisms With Machine Learning Approaches, Front Psychiatry 11, (2020).

[7] Tobacco Control, https://www.who.int/data/gho/data/themes/theme-details/GHO/tobacco-control.

[8] N. A. Rigotti, G. R. Kruse, J. Livingstone-Banks, and J. Hartmann-Boyce, Treatment of Tobacco Smoking: A Review, JAMA 327, 566 (2022).

[9] G. Yang, Y. Wang, Y. Wu, J. Yang, and X. Wan, The Road to Effective Tobacco Control in China, Lancet 385, 1019 (2015).

[10] Y. Ma et al., Prevalence of Cigarette Smoking and Nicotine Dependence in Men and Women Residing in Two Provinces in China, Front Psychiatry 8, (2017).
[11] J. Koplan and M. Eriksen, Smoking Cessation for Chinese Men and Prevention for Women, Lancet 386, 1422 (2015).
[12] C. Zhu, S. Young-Soo, and R. Beaglehole, Tobacco Control in China: Small Steps towards a Giant Leap, Lancet 379, 779 (2012).
[13] C. D. Patnode, J. T. Henderson, J. H. Thompson, C. A. Senger, S. P. Fortmann, and E. P. Whitlock, Behavioral Counseling and Pharmacotherapy Interventions for Tobacco Cessation in Adults, Including Pregnant Women: A Review of Reviews for the U.S. Preventive Services Task Force, Https://Doi.Org/10.7326/M15-0171 163, 608 (2015).
[14] Y. Ren, D. Wu, A. Singh, E. Kasson, M. Huang, and P. Cavazos-Rehg, Automated Detection of Vaping-Related Tweets on Twitter During the 2019 EVALI Outbreak Using Machine Learning Classification, Front Big Data 5, 5 (2022).
[15] N. Shah, M. Nali, C. Bardier, J. Li, J. Maroulis, R. Cuomo, and T. K. Mackey, Applying Topic Modelling and Qualitative Content Analysis to Identify and Characterise ENDS Product Promotion and Sales on Instagram, Tob Control tobaccocontrol (2021).
[16] M. A. Orellana-Barrios, D. Payne, Z. Mulkey, and K. Nugent, Electronic Cigarettes - A Narrative Review for Clinicians, American Journal of Medicine 128, 674 (2015).
[17] J. Drope, Z. Cahn, R. Kennedy, A. C. Liber, M. Stoklosa, R. Henson, C. E. Douglas, and J. Drope, Key Issues Surrounding the Health Impacts of Electronic Nicotine Delivery Systems (ENDS) and Other Sources of Nicotine, CA Cancer J Clin 67, 449 (2017).
[18] J. K. Pepper and N. T. Brewer, Electronic Nicotine Delivery System (Electronic Cigarette) Awareness, Use, Reactions and Beliefs: A Systematic Review, Tob Control 23, 375 (2014).
[19] M. A. Rahman, N. Hann, A. Wilson, and L. Worrall-Carter, Electronic Cigarettes: Patterns of Use, Health Effects, Use in Smoking Cessation and Regulatory Issues, Tob Induc Dis 12, (2014).
[20] Journal of Addiction Medicine : The Official Journal of the American Society of Addiction Medicine. (Journal, Magazine, 2007) [WorldCat.Org], https://www.worldcat.org/title/journal-of-addiction-medicine-the-official-journal-of-the-american-society-of-addiction-medicine/oclc/69420533.
[21] M. A. Orellana-Barrios, D. Payne, Z. Mulkey, and K. Nugent, Electronic Cigarettes - A Narrative Review for Clinicians, American Journal of Medicine 128, 674 (2015).
[22] T. Cheng, Chemical Evaluation of Electronic Cigarettes, Tob Control 23, ii11 (2014).
[23] J. D. Tyzack and J. Kirchmair, Computational Methods and Tools to Predict Cytochrome P450 Metabolism for Drug Discovery, Chemiol Drug Des 93, 377 (2019).
[24] E. E. Litsa, P. Das, and L. E. Kavraki, Machine Learning Models in the Prediction of Drug Metabolism: Challenges and Future Perspectives, Expert Opin Drug Metab Toxicol 17, 1245 (2021).
[25] A. S. Hatoum, F. R. Wendt, M. Galimberti, R. Polimanti, B. Neale, H. R. Kranzler, J. Gelernter, H. J. Edenberg, and A. Agrawal, Ancestry May Confound Genetic Machine Learning: Candidate-Gene Prediction of Opioid Use Disorder as an Example, Drug Alcohol Depend 229, 109115 (2021).
[26] A.Géron, Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media 851 (2019).
[27] A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM J Res Dev 3, 210 (1959).
[28] Master Machine Learning Algorithms, https://machinelearningmastery.com/master-machine-learning-algorithms/.
[29] Practical Machine Learning Problems, https://machinelearningmastery.com/practical-machine-learning-problems/?utm_source=drip&utm_medium=email&utm_campaign=Machine+Learning+Mastery+Crash+Course&utm_content=Practical+machine+learning+problems.
[30] A. Burkov, Machine Learning Engineering (2020).
[31] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang, An Introduction to Machine Learning, Clin Pharmacol Ther 107, 871 (2020).
[32] D. H. Han, S. H. Lee, S. Lee, and D. C. Seo, Identifying Emerging Predictors for Adolescent Electronic Nicotine Delivery Systems Use: A Machine Learning Analysis of the Population Assessment of Tobacco and Health Study, Prev Med (Baltim) 145, (2021).
[33] R. Fu et al., Machine Learning Applications in Tobacco Research: A Scoping Review, Tob Control tobaccocontrol (2021).
[34] R. Chen and J. Lin, Identification of Feature Risk Pathways of Smoking-Induced Lung Cancer Based on SVM, PLoS One 15, e0233445 (2020).
[35] A. M. Chapman, K. Y. Sun, P. Ruestow, D. M. Cowan, and A. K. Madl, Lung Cancer Mutation Profile of EGFR, ALK, and KRAS: Meta-Analysis and Comparison of Never and Ever Smokers, Lung Cancer 102, 122 (2016).
[36] T. Remen, J. Pintos, M. Abrahamowicz, and J. Siemiatycki, Risk of Lung Cancer in Relation to Various Metrics of Smoking History: A Case-Control Study in Montreal 11 Medical and Health Sciences 1117 Public Health and Health Services, BMC Cancer 18, 1 (2018).
[37] R. A. Pleasants, M. P. Rivera, S. L. Tilley, and S. P. Bhatt, Both Duration and Pack-Years of Tobacco Smoking Should Be Used for Clinical Practice and Research, Ann Am Thorac Soc 17, 804 (2020).
[38] K. Davagdorj, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, XGBoost-Based Framework for Smoking-Induced Noncommunicable Disease Prediction, Int J Environ Res Public Health 17, 1 (2020).
[39] S. Kathirvel and J. S. T. Rapporteurs, Sustainable Development Goals and Noncommunicable Diseases: Roadmap till 2030 – A Plenary Session of World Noncommunicable Diseases Congress 2017, Int J Noncommun Dis 3, 3 (2018).
[40] C. B. Breckenridge, C. Berry, E. T. Chang, R. L. Sielken, and J. S. Mandel, Association between Parkinson's Disease and Cigarette Smoking, Rural Living, Well-Water Consumption, Farming and Pesticide Use: Systematic Review and Meta-Analysis, PLoS One 11, e0151841 (2016).
[41] K. Kondo et al., The Association between Environmental Factors and the Development of Crohn's Disease with Focusing on Passive Smoking: A Multicenter Case-Control Study in Japan, PLoS One 14, e0216429 (2019).
[42] S. Li, B. Chen, H. Chen, Z. Hua, Y. Shao, H. Yin, and J. Wang, Analysis of Potential Genetic Biomarkers and Molecular Mechanism of Smoking-Related Postmenopausal Osteoporosis Using Weighted Gene Co-Expression Network Analysis and Machine Learning, PLoS One 16, e0257343 (2021).
[43] J. A. Kanis, E. v. McCloskey, H. Johansson, A. Oden, L. J. Melton, and N. Khaltaev, A Reference Standard for the Description of Osteoporosis, Bone 42, 467 (2008).
[44] A. Taguchi, M. Ohtsuka, T. Nakamoto, K. Naito, M. Tsuda, Y. Kudo, E. Motoyama, Y. Suei, and K. Tanimoto, Identification of Post-Menopausal Women at Risk of Osteoporosis by Trained General Dental Practitioners Using Panoramic Radiographs, Http://Dx.Doi.Org/10.1259/Dmfr/31116116 36, 149 (2014).
[45] P. Xiao et al., In Vivo Genome-Wide Expression Study on Human Circulating B Cells Suggests a Novel ESR1 and MAPK3 Network for Postmenopausal Osteoporosis, Journal of Bone and Mineral Research 23, 644 (2008).
[46] K. K. Venkat, M. M. Arora, P. Singh, M. Desai, and I. Khatkhatay, Effect of Alcohol Consumption on Bone Mineral Density and Hormonal Parameters in Physically Active Male Soldiers, Bone 45, 449 (2009).

[47] C. Oncken, S. Allen, M. Litt, A. Kenny, H. Lando, A. Allen, and E. Dornelas, Exercise for Smoking Cessation in Postmenopausal Women: A Randomized, Controlled Trial, Nicotine Tob Res 22, 1587 (2020).
[48] X. Zhang et al., Machine Learning Selected Smoking-Associated DNA Methylation Signatures That Predict HIV Prognosis and Mortality, Clin Epigenetics 10, 1 (2018).
[49] M. Helleberg et al., Smoking and Life Expectancy among HIV-Infected Individuals on Antiretroviral Therapy in Europe and North America, AIDS 29, 221 (2015).
[50] K. v. Ruggles, Y. Fang, J. Tate, S. M. Mentor, K. J. Bryant, D. A. Fiellin, A. C. Justice, and R. S. Braithwaite, What Are the Patterns Between Depression, Smoking, Unhealthy Alcohol Use, and Other Substance Use Among Individuals Receiving Medical Care? A Longitudinal Study of 5479 Participants, AIDS Behav 21, 2014 (2017).
[51] F. Marabita et al., Smoking Induces DNA Methylation Changes in Multiple Sclerosis Patients with Exposure-Response Relationship, Scientific Reports 2017 7:1 7, 1 (2017).
[52] F. Fasanelli et al., Hypomethylation of Smoking-Related Genes Is Associated with Future Lung Cancer in Four Prospective Cohorts, Nature Communications 2015 6:1 6, 1 (2015).
[53] X. Gao, X. Gào, Y. Zhang, L. P. Breitling, B. Schöttker, and H. Brenner, Associations of Self-Reported Smoking, Cotinine Levels and Epigenetic Smoking Indicators with Oxidative Stress among Older Adults: A Population-Based Study, European Journal of Epidemiology 2017 32:5 32, 443 (2017).
[54] R. Philibert et al., Reversion of AHRR Demethylation Is a Quantitative Biomarker of Smoking Cessation, Front Psychiatry 7, 55 (2016).
[55] Y. Zhang, I. Florath, K. U. Saum, and H. Brenner, Self-Reported Smoking, Serum Cotinine, and Blood DNA Methylation, Environ Res 146, 395 (2016).
[56] X. Gao, M. Jia, Y. Zhang, L. P. Breitling, and H. Brenner, DNA Methylation Changes of Whole Blood Cells in Response to Active Smoking Exposure in Adults: A Systematic Review of DNA Methylation Studies, Clin Epigenetics 7, 1 (2015).
[57] M. J. Corley et al., Comparative DNA Methylation Profiling Reveals an Immunoepigenetic Signature of HIV-Related Cognitive Impairment, Scientific Reports 2016 6:1 6, 1 (2016).
[58] X. Zhang, Y. Hu, A. C. Justice, B. Li, Z. Wang, H. Zhao, J. H. Krystal, and K. Xu, DNA Methylation Signatures of Illicit Drug Injection and Hepatitis C Are Associated with HIV Frailty, Nature Communications 2017 8:1 8, 1 (2017).
[59] S. Horvath and A. J. Levine, HIV-1 Infection Accelerates Age According to the Epigenetic Clock, J Infect Dis 212, 1563 (2015).
[60] K. N. Nelson, Q. Hui, D. Rimland, K. Xu, M. S. Freiberg, A. C. Justice, V. C. Marconi, and Y. v. Sun, Identification of HIV Infection-Related DNA Methylation Sites and Advanced Epigenetic Aging in HIV-Positive, Treatment-Naive U.S. Veterans, AIDS 31, 571 (2017).
[61] P. M. Dietz, L. J. England, C. K. Shapiro-Mendoza, V. T. Tong, S. L. Farr, and W. M. Callaghan, Infant Morbidity and Mortality Attributable to Prenatal Smoking in the U.S, Am J Prev Med 39, 45 (2010).
[62] C. A. Markunas, Z. Xu, S. Harlid, P. A. Wade, R. T. Lie, J. A. Taylor, and A. J. Wilcox, Identification of DNA Methylation Changes in Newborns Related to Maternal Smoking during Pregnancy, Environ Health Perspect 122, 1147 (2014).
[63] S. H. Timmermans, M. Mommers, J. S. Gubbels, S. P. J. Kremers, A. Stafleu, C. D. A. Stehouwer, M. H. Prins, J. Penders, and C. Thijs, Maternal Smoking during Pregnancy and Childhood Overweight and Fat Distribution: The KOALA Birth Cohort Study, Pediatr Obes 9, e14 (2014).
[64] E. Oken, E. B. Levitan, and M. W. Gillman, Maternal Smoking during Pregnancy and Child Overweight: Systematic Review and Meta-Analysis, International Journal of Obesity 2008 32:2 32, 201 (2007).
[65] P. Wiklund et al., DNA Methylation Links Prenatal Smoking Exposure to Later Life Health Outcomes in Offspring, Clin Epigenetics 11, 1 (2019).
[66] S. Rauschert et al., Machine Learning-Based DNA Methylation Score for Fetal Exposure to Maternal Smoking: Development and Validation in Samples Collected from Adolescents and Adults, Environ Health Perspect 128, 1 (2020).
[67] M. V. C. Greenberg and D. Bourc'his, The Diverse Roles of DNA Methylation in Mammalian Development and Disease, Nature Reviews Molecular Cell Biology 2019 20:10 20, 590 (2019).
[68] Y. v. Sun et al., Epigenomic Association Analysis Identifies Smoking-Related DNA Methylation Sites in African Americans, Human Genetics 2013 132:9 132, 1027 (2013).
[69] [P. Rzehak et al., Maternal Smoking during Pregnancy and DNA-Methylation in Children at Age 5.5 Years: Epigenome-Wide-Analysis in the European Childhood Obesity Project (CHOP)-Study, PLoS One 11, e0155554 (2016).
[70] B. R. Joubert et al., DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-Wide Consortium Meta-Analysis, Am J Hum Genet 98, 680 (2016).
[71] W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, (2002).
[72] J. Parks et al., Assessing Secondhand and Thirdhand Tobacco Smoke Exposure in Canadian Infants Using Questionnaires, Biomarkers, and Machine Learning, Journal of Exposure Science & Environmental Epidemiology 2021 32:1 32, 112 (2021).
[73] J. Yang, S. Wang, Z. Yang, C. A. Hodgkinson, P. Iarikova, J. Z. Ma, T. J. Payne, D. Goldman, and M. D. Li, The Contribution of Rare and Common Variants in 30 Genes to Risk Nicotine Dependence, Molecular Psychiatry 2014 20:11 20, 1467 (2014).
[74] A. Jamal, E. Phillips, A. S. Gentzke, D. M. Homa, S. D. Babb, B. A. King, and L. J. Neff, Current Cigarette Smoking Among Adults — United States, 2016, MMWR Morb Mortal Wkly Rep 67, 53 (2022).
[75] S. Zhou, D. G. Rosenthal, S. Sherman, J. Zelikoff, T. Gordon, and M. Weitzman, Physical, Behavioral, and Cognitive Effects of Prenatal Tobacco and Postnatal Secondhand Smoke Exposure, Curr Probl Pediatr Adolesc Health Care 44, 219 (2014).
[76] P. Jacob et al., Thirdhand Smoke: New Evidence, Challenges, and Future Directions, Chem Res Toxicol 30, 270 (2017).
[77] E-Cigarettes, Vapes, and Other Electronic Nicotine Delivery Systems (ENDS) | FDA, https://www.fda.gov/tobacco-products/products-ingredients-components/e-cigarettes-vapes-and-other-electronic-nicotine-delivery-systems-ends.
[78] Vaping and E-Cigarettes: A Toolkit for Working With Youth - Tobacco Education Resource Library Print Materials & Downloads, https://digitalmedia.hhs.gov/tobacco/print_materials/CTP-218?locale=en.
[79] Center for Tobacco Products | FDA, https://www.fda.gov/about-fda/fda-organization/center-tobacco-products.
[80] J. Drope, Z. Cahn, R. Kennedy, A. C. Liber, M. Stoklosa, R. Henson, C. E. Douglas, and J. Drope, Key Issues Surrounding the Health Impacts of Electronic Nicotine Delivery Systems (ENDS) and Other Sources of Nicotine, CA Cancer J Clin 67, 449 (2017).
[81] J. K. Pepper and N. T. Brewer, Electronic Nicotine Delivery System (Electronic Cigarette) Awareness, Use, Reactions and Beliefs: A Systematic Review, Tob Control 23, 375 (2014).
[82] M. A. Orellana-Barrios, D. Payne, Z. Mulkey, and K. Nugent, Electronic Cigarettes - A Narrative Review for Clinicians, American Journal of Medicine 128, 674 (2015).
[83] Results from the Annual National Youth Tobacco Survey | FDA, https://www.fda.gov/tobacco-products/youth-and-tobacco/results-annual-national-youth-tobacco-survey.
[84] C. O. on S. and Health, Smoking and Tobacco Use; Data and Statistics; Surveys; National Youth Tobacco Survey (NYTS), (2022).
[85] C. O. on S. and Health, Smoking and Tobacco Use; Data and Statistics; Surveys; National Youth Tobacco Survey (NYTS), (2023).

[86] J. Wagner, W. Chen, and G. Vrdoljak, Vaping Cartridge Heating Element Compositions and Evidence of High Temperatures, PLoS One 15, (2020).
[87] X. Zeng, X. Xu, H. M. Boezen, and X. Huo, Children with Health Impairments by Heavy Metals in an E-Waste Recycling Area, Chemosphere 148, 408 (2016).
[88] K. Jomova and M. Valko, Advances in Metal-Induced Oxidative Stress and Human Disease, Toxicology 283, 65 (2011).
[89] A. Alavalapadu and R. Mattamal, Vaping Associated Pulmonary Injury, Int J Integr Pediatr Environ Med 7, 8 (2022).
[90] B. C. Blount et al., Smoking and Tobacco Use; Electronic Cigarettes, New England Journal of Medicine 382, 697 (2021).
[91] G. S. Casanova, R. Amaro, N. Soler, M. Sánchez, J. R. Badía, J. A. Barberà, and A. Agustí, An Imported Case of E-Cigarette or Vaping Associated Lung Injury in Barcelona, Eur Respir J 55, (2020).
[92] C. Marlière, J. de Greef, S. Gohy, D. Hoton, P. Wallemacq, L. M. Jacquet, and L. Belkhir, Fatal E-Cigarette or Vaping Associated Lung Injury (EVALI): A First Case Report in Europe, Eur Respir J 56, (2020).
[93] I. Barrientos-Gutierrez, P. Lozano, E. Arillo-Santillan, P. Morello, R. Mejia, and J. F. Thrasher, "Technophilia": A New Risk Factor for Electronic Cigarette Use among Early Adolescents?, Addictive Behaviors 91, 193 (2019).
[94] S. Lee, D. H. Han, A. Chow, and D. C. Seo, A Prospective Longitudinal Relation between Elevated Use of Electronic Devices and Use of Electronic Nicotine Delivery Systems, Addictive Behaviors 98, 106063 (2019).
[95] H. J. Kan, H. Kharrazi, H. Y. Chang, D. Bodycombe, K. Lemke, and J. P. Weiner, Exploring the Use of Machine Learning for Risk Adjustment: A Comparison of Standard and Penalized Linear Regression Models in Predicting Health Care Costs in Older Adults, PLoS One 14, e0213258 (2019).
[96] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, and G. R. Bernard, Identifying Tweets of Personal Health Experience through Word Embedding and LSTM Neural Network, BMC Bioinformatics 19, 67 (2018).
[97] R. Skaik and Di. Inkpen, Using Social Media for Mental Health Surveillance, ACM Computing Surveys (CSUR) 53, (2020).
[98] D. M. Kazemi, B. Borsari, M. J. Levine, and B. Dooley, Systematic Review of Surveillance by Social Media Platforms for Illicit Drug Use, J Public Health (Oxf) 39, 763 (2017).
[99] Y. Ren, D. Wu, A. Singh, E. Kasson, M. Huang, and P. Cavazos-Rehg, Automated Detection of Vaping-Related Tweets on Twitter During the 2019 EVALI Outbreak Using Machine Learning Classification, Front Big Data 5, 5 (2022).
[100] C. C. Xavier and M. Souza, A Basic Approach for Extracting and Analyzing Data from Twitter, Special Topics in Multimedia, IoT and Web Technologies 185 (2020).
[101] How to Scrape Tweets from Twitter with Python Twint | by Andika Pratama | Analytics Vidhya | Medium, https://medium.com/analytics-vidhya/how-to-scrape-tweets-from-twitter-with-python-twint-83b4c70c5536.
[102] G. Kostygina, H. Tran, L. Czaplicki, S. N. Perks, D. Vallone, S. L. Emery, and E. C. Hair, Developing a Theoretical Marketing Framework to Analyse JUUL and Compatible E-Cigarette Product Promotion on Instagram, Tob Control 0, tobaccocontrol (2022).
[103] M. Hu, R. Benson, A. T. Chen, S. H. Zhu, and M. Conway, Determining the Prevalence of Cannabis, Tobacco, and Vaping Device Mentions in Online Communities Using Natural Language Processing, Drug Alcohol Depend 228, (2021).
[104] R. Fu et al., Machine Learning Applications in Tobacco Research: A Scoping Review, Tob Control tobaccocontrol (2021).