Toxicology Reports 14 (2025) 101877

Contents lists available at ScienceDirect

Toxicology Reports

journal homepage: www.elsevier.com/locate/toxrep

Harnessing machine learning in contemporary tobacco research

Krishnendu Sinha [a,*], Nabanita Ghosh [b], Parames C. Sil [c,*]

[a] Jhargram Raj College, Jhargram 721507, India
[b] Maulana Azad College, Kolkata 700013, India
[c] Division of Molecular Medicine, Bose Institute, Kolkata 700054, India

ARTICLE INFO

Handling Editor: Dr. L.H. Lash

Keywords:
Machine learning
Algorithms
Smoking
Tobacco
Vapes
Cancer

ABSTRACT

Machine learning (ML) has the potential to transform tobacco research and address the urgent public health crisis posed by tobacco use. Despite the well-documented health risks, cessation rates remain low. ML techniques offer innovative solutions by analyzing vast datasets to uncover patterns in smoking behavior, genetic predispositions, and effective cessation strategies. ML can predict smoking-induced non-communicable diseases (SiNCDs) like lung cancer and postmenopausal osteoporosis by identifying biomarkers and genetic profiles, generating personalized predictions, and guiding interventions. It also improves prediction of infant tobacco smoke exposure, distinguishes secondhand and thirdhand smoke, and enhances protection strategies for children. Data-driven, personalized approaches using ML track real-time data for personalized feedback and offer timely interventions, continuously improving cessation strategies. Overall, ML provides sophisticated predictive models, enhances understanding of complex biological mechanisms, and enables personalized interventions, demonstrating significant potential in the fight against the tobacco epidemic.

1. Introduction

Tobacco smoking poses a significant public health threat due to its immediate and long-term detrimental effects on physical health, including increased risk of respiratory infections, various cancers, cardiovascular diseases, and weakened immune function [1–4]. In 2020, the World Health Organization (WHO) reported that 36.7 % of men and 7.8 % of women, totalling 22.3 % of the global population, used tobacco [5]. Tobacco use is the leading preventable cause of death worldwide, causing 8 million deaths annually, including 1.2 million non-smokers affected by passive smoking [2,6]. Without effective tobacco control efforts, this number is projected to increase to 18.3 million by 2030 [7]. In the U.S., approximately 14 % of adults, or 34 million people, are active combustible tobacco users, who could gain up to 10 years of life by quitting [8]. Prevention is critical in addressing the tobacco use pandemic [9]. Around 70 % of people who use combustible tobacco want to quit but typically need an average of 6 attempts to achieve lasting abstinence, with nicotine replacement therapy (NRT) and counselling being effective [8]. Tobacco use is a complex behavior influenced by genetic factors (up to 50 %) and various socio-environmental factors like peer influence, workplace culture, social gatherings, media and advertising, stressful situations, community regulations, presence of triggers etc [8]. Additionally, electronic nicotine delivery systems (ENDS), such as e-cigarettes, are becoming a public health concern, particularly among adolescents. Although ENDS were initially developed to help people quit tobacco, their rapid and unregulated growth has posed significant risks to public health, leading to increased scrutiny and regulation, especially regarding advertising and sales on social media [10].

In recent years, the shift from traditional computational methods to machine learning (ML) models has been driven by the availability of vast amounts of data [11,12]. ML is now making significant impacts in fields such as psychology, medicine, and public health [12]. Researchers are using ML to tackle complex issues in tobacco research, such as understanding psycho-genetic predispositions to tobacco addiction, analyzing intricate behavioral patterns in people who use tobacco, providing personalized digital tools to assist in quitting, and monitoring unregulated vaping trades on social media [12,13]. These advancements, once considered unimaginable, are now becoming integral in the fight against the tobacco epidemic, highlighting the transformative potential of next-generation ML algorithms and big data research [14].

In this review, we aim to formulate four comprehensive questions (Table 1) shaping contemporary tobacco research and identify the limitations of traditional methodologies. By exploring the transformative

* Corresponding authors.
E-mail addresses: dr.krishnendusinha@gmail.com (K. Sinha), parames@jcbose.ac.in (P.C. Sil).

https://doi.org/10.1016/j.toxrep.2024.101877
Received 4 September 2024; Received in revised form 17 December 2024; Accepted 17 December 2024
Available online 19 December 2024

potential of ML, we demonstrate how it can effectively address these challenges and fill existing research gaps. Our goal is to provide researchers with a comprehensive, up-to-date overview of ML applications in tobacco research, empowering them to leverage these insights in their work (Fig. 1; Table 2). We believe this review will serve as a valuable resource, fostering innovation and collaboration within the research community.

2. Literature search strategy

The literature search for this review employed a methodical approach to comprehensively cover advancements in ML-assisted tobacco research. Searches were conducted across databases including PubMed, Scopus, and Web of Science, using keywords like "machine learning", "ML", "tobacco research", "smoking", "tobacco", "vaping", "vape" and related terms. Boolean operators (AND, OR) were utilized to refine search queries and ensure inclusivity of relevant literature from January 2010 to June 2024. Additionally, hand-searching of reference lists and inclusion of grey literature such as conference proceedings supplemented the search. Criteria for inclusion encompassed studies employing ML techniques for analyzing tobacco-related data, including predictive modeling, pattern recognition, and sentiment analysis, focusing on tobacco use, cessation interventions, health outcomes, and policy evaluations. This methodical approach aimed to provide a comprehensive overview aligned with current research trends in ML applications within tobacco research.

3. A Brief account on ML

ML is the science of programming computer so that the computer can learn from data [15]. Conventional programming works by setting a directory of certain rules, which a computer follows to provide desirable outcome. In other words, the information goes in, rules applied, and the results come up. This approach turns out to be ridiculously challenging in the case of intricate dynamic situations like, assessing human emotions to addiction, extracting predictor for adolescent nicotine addiction, etc. On contrary, ML does not require setting of explicit rules to produce result. Through ML algorithms, computers extract pattern from input data and set instance specific rules. Historically ML is defined as the field of study that empowers computers to learn without being explicitly programmed [16]. In general, most of the complex ML challenges can be reduced to one of the four core problem types namely, Classification, Regression, Clustering and Rule extraction [17,18]. Classification deals with labelling of discreate data points, meaning it assigned a class (e.g., smoker/non-smoker or ever-smoker/never-smoker) to a data point [19]. Based on these labelled datasets, a classification model gets train to allocate labels to new unlabeled data points. This can be understood as a discrimination problem, demonstrating the disparities or similarities among groups [18]. Regression deals with continuous numerical data points (generally floats). A regression model after being trained, can predict numerical outcome for new unpredicted data, like the average age of start smoking in a population under study [19,20]. In clustering, unlabeled data can be split into groups based on similarity and additional measures of the inherent natural data structure. Identification of e-cigarettes related tweets from other non-specific tweets can be considered as clustering problem. Lastly, in the case of rule extraction problem, data is utilized for the extraction of propositional laws. Such rules are not usually directed, that mean, these methods acquire statistically acceptable associations among properties in the data, not essentially involving something that is being predicted [18]. An instance is the finding of the connection among the addiction of e-cigarettes, marijuana, or alcohol

Table 1
Key research questions in tobacco control, associated problem types, applicable machine learning models, and potential outcomes.

Research Question	Problem Type	ML Model Type	Model Description	Examples of Previous Work	Potential Outcomes
How can we predict the individual health impacts of smoking?	Personalized Predictions, Risk Assessment, Classification, Predictive Analytics	Random Forest (RF), Support Vector Machine (SVM), Neural Networks (NN), Gradient Boosting Machines (GBM)	RF: Ensemble of decision trees for classification, SVM: Hyperplane classification, NN: Complex pattern recognition, GBM: Sequential model building	[14–33]	Accurate individual health risk assessments, personalized intervention strategies, improved health outcomes
How can we accurately assess and monitor passive smoke exposure?	Risk Assessment, Environmental Monitoring, Predictive Analytics, Anomaly Detection	Logistic Regression, Decision Trees, Bayesian Networks, Gradient Boosting Machines (GBM), Random Forest, Neural Networks, Support Vector Machine (SVM)	Logistic Regression: Probability estimation, Decision Trees: Simple classification, Bayesian Networks: Probabilistic graphical models, GBM: Sequential model building, RF: Ensemble of trees, NN: Complex pattern recognition, SVM: Hyperplane classification	[34–39]	Proactive protection measures, real-time exposure monitoring, tailored interventions, improved monitoring of exposure levels, better public health strategies
How to predict and improve smoking cessation outcome?	Prediction, Classification, Causal Inference	Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, Neural Networks (NN)	RF: Ensemble of decision trees, SVM: Hyperplane classification, LR: Probability prediction, NN: Complex patterns recognition	[40–60]	Improved prediction models for intervention effectiveness, targeted cessation programs, higher cessation rates, tailored support programs, reduced relapse rates
What is the impact of evolving tobacco product landscape on youth and how to mould intervention strategies accordingly?	Behavioral Analysis, Pattern Recognition, Clustering, Time-Series Analysis, Social Media Analysis, Natural Language Processing, Impact Assessment, Intervention Design	K-Means Clustering, Neural Networks (NN), Hidden Markov Models (HMM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Transformers, Natural Language Processing (NLP), Clustering Algorithms	K-Means: Data partitioning, NN: Complex pattern recognition, HMM: State transitions, RNN: Sequence modeling, CNN: Image data processing, LSTM: Sequence data processing, Transformers: Advanced NLP models, NLP: Understand, interpret, and manipulate human language, Clustering: Identifying youth segments based on behavior	[61–83]	Insights into youth behavioral patterns, development of effective youth-focused interventions, better regulation and tracking of vaping products, reduced initiation rates

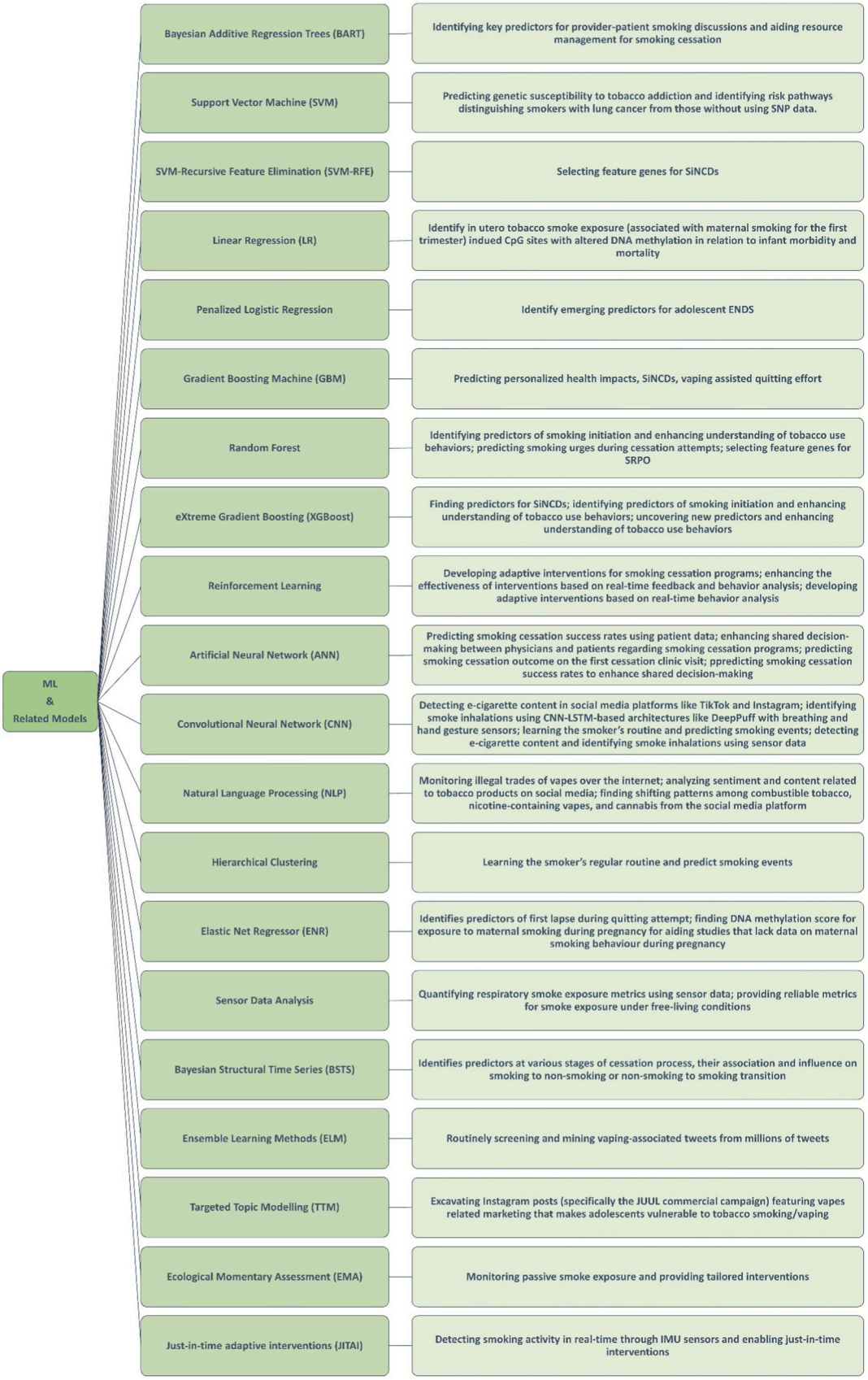

Fig. 1. A hierarchical chart illustrating how different ML approaches have been used to solve various tobacco research problems.

Table 2
Comprehensive glossary.

Term	Definition
Adaptive Boosting (AdaBoost)	A machine learning meta-algorithm that combines weak learners to create a strong learner, used for classification and regression tasks.
Artificial Neural Networks (ANN)	A machine learning model inspired by the structure and function of biological neural networks, used for various tasks such as classification, regression, and prediction.
Bayesian Additive Regression Trees (BART)	A non-parametric Bayesian approach to modeling data, used in tobacco research for recognizing key determinants promoting smoking cessation discussions.
Convolutional Neural Networks (CNN)	A type of artificial neural network used in image recognition and processing, designed to automatically and adaptively learn spatial hierarchies of features.
Decision Tree (DT)	A supervised machine learning algorithm used for classification and regression tasks.
Deep Learning	A subset of machine learning involving neural networks with many layers (deep neural networks) that learn from large amounts of data. Includes models like CNNs for visual analysis and LSTMs for sequential data processing.
DeepPuff	A CNN-LSTM-based deep learning architecture quantifying Respiratory Smoke Exposure Metrics (RSEM) by detecting smoke inhalations via sensors.
DNA Methylation	A process by which methyl groups are added to DNA molecules, affecting gene expression, used in ML models to identify fetal exposure to maternal smoking.
Ecological Momentary Assessment (EMA)	A research method that involves collecting real-time data on participants' behaviors, symptoms, and contexts in their natural environments.
Electronic Nicotine Delivery System (ENDS)	Also known as e-cigarettes, they are becoming a public health concern, particularly among adolescents, despite being initially developed to help people quit tobacco.
eXtreme Gradient Boosting (XGBoost)	A scalable and efficient implementation of gradient boosting, a machine learning technique that produces a prediction model in the form of an ensemble of weak prediction models.
Gradient Boosting Machine (GBM)	A machine learning technique that produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
Hierarchical Cluster Analysis (HCA)	An unsupervised machine learning algorithm used for clustering data.
Just-in-time adaptive interventions (JITAI)	Interventions that aim to provide the right type/amount of support at the right time by adapting to an individual's changing internal and contextual state.
k-Nearest Neighbors (KNN)	A non-parametric method used for classification and regression, where the output is a class membership or a property value for an object based on the k nearest training examples in the feature space.
Least Absolute Shrinkage and Selection Operator (LASSO)	A regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
Light Gradient Boosting Machine (LGBM)	A highly efficient gradient boosting framework using tree-based learning algorithms, applied to pinpoint predictors of SHS-induced depression.
Linear Regression (LR)	A statistical method used to model the relationship between a dependent variable and one or more independent variables using a straight line.
Machine Learning (ML)	A shift from traditional computational methods driven by the availability of vast amounts of data, making significant impacts in fields such as psychology, medicine, and public health to tackle complex issues in tobacco research.
Nicotine Replacement Therapy (NRT)	Along with provider-patient discussion, is recommended for treating nicotine dependence, but the rate of success is not encouraging.
Principal Component Analysis (PCA)	A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
Random Forest (RF)	An ensemble learning method for classification, regression and other tasks, operating by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.
Secondhand Smoke (SHS)	Smoke that non-smokers involuntarily inhale from nearby smokers, posing significant public health risks linked to respiratory infections, cardiovascular diseases, and cancer.
SHapley Additive explanation (SHAP)	A method for explaining ML model outputs by attributing each feature's contribution to predictions, used in assessing smoking cessation predictors.
Single Nucleotide Polymorphism (SNP)	A variation in a single nucleotide at a specific genome position, used in ML models to predict smoking behavior.
Smoking Induced Non-Communicable Diseases (SiNCDs)	Include stroke, heart disease, chronic respiratory diseases, cancers, etc., which can have significantly reduced death rates through early diagnosis, effective treatment, and smoking cessation.
Support Vector Machine (SVM)	A supervised machine learning model that analyzes data for classification and regression analysis.
Support Vector Machine Recursive Feature Elimination (SVM-RFE)	A feature selection algorithm that uses a SVM as the classifier and recursively removes features based on their weights until the desired number of features is reached.
t-Distributed Stochastic Neighbor Embedding (t-SNE)	A dimensionality reduction technique used for visualizing high-dimensional data.
Thirdhand Smoke (THS)	Residual contaminants that linger on surfaces and dust, posing health risks, especially to children.
Tobacco smoking	A highly complex behavior controlled by genetic predisposition and environmental parameters, posing serious health risks and being a leading preventable cause of death worldwide.
Weighted Gene Co-expression Network Analysis (WGCNA)	A systems biology method used to describe the correlation patterns among genes across microarray samples, with the goal of finding clusters (modules) of highly correlated genes, and relating these modules to external sample traits.

[19]. Commonly a vast majority of ML systems broadly fall under any of these four categories namely, supervised, unsupervised, semi-supervised, and reinforcement learning based on the degree of human supervision involved in model training to solve the four core problem types [15,21,22].

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 each element x_i represents a feature vector, and the dimensionality of the dataset corresponds to the length of this vector [21]. 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, individual row signifying a vector for one instance and individual column signifying all the instances' values for that feature. As 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 [19,21]. 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 [15,21]. Decision Trees (DTs) and Random Forests (RFs), *k*-Nearest Neighbors (*K*NN), Support Vector Machines (SVMs), Linear Regression (LR), artificial neural networks (ANN) are some contextually relevant supervised learning algorithms. However, cumulatively it can be stated that the intent of a supervised learning algorithm is to utilize a labeled dataset to generate a model which in succession takes an unlabeled feature vector x_i as input and outputs a label for that [21].

Unsupervised learning works with a set of unlabeled examples called training dataset. Mathematically speaking, the training dataset is represented as $\{x_i\}_{i=1}^{n}$, where every element x_i is a feature vector. On contrary to supervised learning, unsupervised learning algorithm produce 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 utilized to resolve a practical problem [15,21]. 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 label/ID to unlabeled data point based on its features[21]. Hierarchical Cluster Analysis (HCA), *K*-Means, Isolation Forest, t-Distributed Stochastic Neighbor Embedding (t-SNE), Principal Component Analysis (PCA) are few vital unsupervised learning clustering algorithms[15]. Examples of unsupervised learning in tobacco research contain discovering the topics of tobacco-related conversations on social media or finding possible subtypes of nicotine addiction by evaluating the brain MRI data of patients[23].

In the following sub-sections, a few widely adopted ML algorithms in the field of tobacco research are briefly discussed.

3.1. Naïve Bayes

Naïve Bayes is a probabilistic ML algorithms built on Bayes Theorem and largely used as a simple and powerful classifier, which assigns the maximum possible class of every single data point, matching to the description provided by the feature vector and assuming the independence of the predictors from each other [24–27]. However, Naïve Bayes obtain remarkable results even under the situation of strong dependencies between predictors and works beautifully with small noisy datasets [25,28].

3.2. Logistic regression

Logistic regressor (LR) can act as a binary classifier, which evaluates a data point's probability of fitting into a particular class where a probability greater than 0.5 is considered as positive belonging to a particular class under observation [15]. Logistic regression model subtracts a weighted sum of the input features in addition to a bias term and, outputs the results [15]. A logistic sigmoid function outputs any float between 0 and 1. It is a very popular algorithm in ML studies.

3.3. k-Nearest Neighbours

k-Nearest Neighbours (*K*NN) is a supervised, non-parametric ML algorithm that identifies the closest neighbors to a given query point to determine its label or classification [29–31]. It is mostly used as a classifier, though it can perform a regression job. *K*NN presumes that, similar data points reside in vicinity. Distances between the query point and rest data points is calculated to infer the closest data points, and these distance metrics assist to form decision boundaries. These boundaries are commonly envisioned by Voronoi tessellation. Decision boundaries categorise query points in distinct classes. The *k* value is a critical hyperparameter that defines the number of neighbours to be checked for classifying a query point. As *K*NN algorithm commonly depend on Euclidian distance for the sake of classification, normalization generally boosts its performance drastically and the algorithm becomes sensitive to local data structure [29,30,32,33]. Since *K*NN is part of a category of algorithms known for their lazy learning approach, its uses in an instance of substantially large dataset becomes prohibitively slow [29].

3.4. Decision trees

Decision Trees (DTs) are flexible non-parametric supervised learning algorithms applied intensively for classification and regression task [15]. Decision tree predicts the value of a target variable by inferring straightforward decision rules from predictors. A decision tree is a hierarchical construct starting at root node representing entire dataset, ending at terminal nodes representing classes resulting from tree analysis, and internal nodes in between, representing decision arguments [25]. Though individual decision trees have little, or no practical use and ensemble of rather uncorrelated trees form a random forest (RF) which is one of the most powerful and widely used ML model. This massive success of the ensemble method stands on a relatively basic perception of the 'wisdom of crowd'. It is well-recognized that the combined responses from thousands of random individuals to a complex question often perform remarkably well, even rivaling the accuracy of an expert's answer [15]. Similarly, collective predictions of a collection of decision trees will often get improved forecasts in contrast to the best isolated predictor [15]. This collection of predictors is termed as ensemble and the process is entitled as the ensemble learning [15,21,25]. Random forest is a legendary ensemble method where, predictors are decision trees. To generate a prediction, the random forest forecasts the class that gets the most votes from individual decision trees [15]. Decision trees are exceptionally sensitive to the dataset for training and minor alteration in the dataset produce significant variation on tree morphology which satisfy the condition of independence between constituting tress of random forest. This is known as bootstrap aggregation or bagging. Bagging, together with few other methods like feature randomness, pasting etc., produce sufficiently random uncorrelated trees in a forest [15,25,33,34].

3.5. Boosting algorithms

Boosting is an ensemble technique that combines multiple weak models, such as linear regression or shallow decision trees, to achieve high-quality predictions by training each model sequentially [35,36]. In this process, each subsequent model addresses the shortcomings of the previous one. Adaptive Boosting (AdaBoost) and Gradient Boosting are two widely recognized boosting techniques, both of which iteratively add weak predictors to the ensemble, with each predictor improving on its predecessor [35]. eXtreme gradient boosting (XGBoost), a popular implementation of gradient-boosted decision trees, is an optimized, distributed gradient boosting library known for its flexibility, efficiency, and portability. It has been extensively utilized in tobacco research [37,38].

3.6. Support vector machines

Support Vector Machine (SVM) is a prevailing, extensively used ML algorithm with the capability of handling high dimensional, non-linear, complex, noisy data and can be used both as classifier and regressor [25]. SVM works by deploying a hyperplane with *N*-1 dimension in an

N-dimensional space (where *N* is the number of features) that explicitly classifies the data points based on their position in respect to the hyperplane. For a two-dimensional data, a hyperplane can be visualized a one-dimensional line that acts a decision boundary. The data points on either side of the boundary can be assigned distinct classes [15,25]. Multiple hyperplanes may exist to separate two classes of data points, but the hyperplane with the maximum margin is considered ideal. Essentially, in the simplest scenario, an SVM classifier can be visualized as creating the widest possible "street" that separates two classes in a two-dimensional plane [15,39]. Adding new data points outside this street does not affect the decision boundary, as it is solely determined by the data points situated at the street's edges [15,40]. These critical data points, referred to as support vectors, define the hyperplane's orientation and position. For complex datasets, such as those used in vape-related sentiment analysis or smoking pattern studies, linear separation is not feasible [23,41,42]. In such cases, SVM employs kernel techniques to handle the complexity. Kernels enable the resolution of non-linear problems by transforming non-linearly separable data into a higher-dimensional space where linear separation becomes possible. Kernel functions achieve this transformation, making the data linearly distinguishable in the new space. Kernel-based methods, including SVM, are highly effective in high-dimensional classification tasks due to their capacity to generalize within such spaces [25,39,40,43].

3.7. Artificial neural networks

Artificial neural network (ANN) is a deep learning algorithm made up of artificial neurons or linear units which simulate the basic architecture of brain in living organism. Linear units are nothing but linear equations which takes input to process with assigned weight along with prefixed bias and gives the output [44]. Neural networks typically unify their linear units into layers and linear units taking a common set of inputs establish a dense layer which by stacking together forms a deep neural network [45]. Despite simple transformation performed by each layer, a deep neural network tries to approximate complex thinking process of brain [45]. Nevertheless, dense layer composed of linear units are not suitable for real world problem as it can never go outside the linearity and thus an activation function like rectified linear unit (ReLU) is added in between two dense layers to import non-linearity into the system and makes stacking effective [44–47].

Inspired by the concept of an artificial neuron, multiple neurons can be connected to form a network, where the output of one neuron serves as the input for another. The input layer receives data from external sources, typically in a vectorized or tensor format, while the output layer provides the final results. In regression tasks, the output layer functions as a simple linear unit, whereas for classification tasks, it incorporates an activation function [25,45]. The intermediate nodes, situated between the input and output layers, are referred to as hidden layers because their outputs are not directly accessible to the user [15,44]. Deep learning represents a specialized form of neural networks, usually consisting of a large number of densely interconnected layers [48]. Its core principle lies in hierarchical information processing, where each layer of neurons works to derive increasingly meaningful representations of the data. Lower layers capture basic features, while subsequent layers combine these simpler elements to model more intricate relationships [25,49].

4. Personalized health impact prediction

Conventional research on tobacco-induced health risks faces significant gaps in the personalized health impact prediction from tobacco use. This could be effectively addressed by harnessing the power of ML, which can analyze vast datasets to identify biomarkers and genetic profiles predisposing individuals to diseases like lung cancer and cardiovascular conditions, generate personalized predictions of adverse smoking outcomes, enhance information extraction from scientific literature, and forecast long-term health outcomes to guide public health interventions and policies [50]. By using ML techniques, researchers identified specific gene isoforms that are significantly different between current and former smokers [51]. These isoforms were identified through a process that included feature selection and ranking algorithms, leading to the development of classification models that can effectively differentiate between the two groups. The identified isoforms and pathways provide insights into the biological impact of smoking, which can help in developing targeted personalized interventions and treatments for smoking-related diseases. ML is also being used to identify specific pathways through which smoking causes damage, especially in the presence of certain comorbid conditions. Toon et al. applied ML to analyze proteomic data from cultured bronchial epithelial cells exposed to cigarette smoke extract, identifying ferroptosis as the most distinctive and significantly affected pathway in COPD patients compared to non-COPD individuals [52]. ML-based feature selection enabled the identification of this key pathway, highlighting the particular vulnerability of COPD epithelial cells to smoke and advancing our understanding of COPD pathogenesis. This could help start personalized treatment plans for COPD patients who smoke.

Non-communicable disease (NCDs) account for 70 % of global deaths, with tobacco as a key cause [53]. Smoking undermines the UN's Sustainable Development Goals, but cessation could reduce the NCD burden by one-third by 2030 [54]. Smoking Induced Non-Communicable Diseases (SiNCDs) such as stroke, heart disease, chronic respiratory diseases [55], cancers, etc., can have significantly reduced death rates through early diagnosis, effective treatment, and smoking cessation [56,57]. ML demonstrates significant potential in predicting various aspects of the likelihood of SiNCDs, which we will discuss in the next section.

4.1. SiNCDs

ML is being utilized to predict the likelihood of developing SiNCDs accurately [58]. The study proposed an effective XGBoost framework integrated with a hybrid feature selection (HFS) technique for SiNCD prediction [58]. This XGBoost framework was applied to real-world NHANES datasets from South Korea and the United States [58]. Comparative analysis showed that the proposed model not only addressed the drawbacks of traditional studies in accurate prediction of SiNCDs in a personalized manner but also outperformed existing baseline models.

4.1.1. Predicting lung cancer

Lung cancer, a major SiNCD, a leading cause of cancer deaths, is strongly linked to prolonged tobacco exposure, with risk increasing with more years of smoking [59]. Most studies focus on the genetic and biological mechanisms connecting smoking to lung cancer. However, many long-term smokers never develop lung cancer, indicating complex interactions and body responses to tobacco [60]. Additionally, lung cancer also occurs in never-smokers, often diagnosed at late stages. Two significant gaps in traditional studies are evident: the need to identify risk pathways of smoking carcinogenesis for early lung cancer diagnosis and personalized treatment in smokers, and the necessity to assess lung cancer risk based on smoking status, especially for never-smokers [61]. Researchers are harnessing ML to address these gaps by providing more sophisticated and accurate risk assessment models [62]. Chen and Lin investigate the impact of smoking factors on lung cancer risk and identifies specific risk pathways associated with smoking-induced lung cancer using ML [60]. They identified five optimal feature pathway sets, which, when combined with clinical information, like gender, age, smoking index, lymphatic lesions, etc, improved diagnostic accuracy to 90 % using a SVM model [60]. The study highlights the potential of the identified pathways as effective risk indicators for differentiating between lung cancer smokers and non-lung cancer smokers, demonstrating their utility in enhancing diagnostic accuracy in clinical settings [60].

Nemlander et al. performed a study which examines how symptoms reported via an adaptive e-questionnaire predict lung cancer across never smokers, along with former and current smokers [61]. Stochastic gradient boosting, stratified by smoking status, was employed to train and test predictive models [61]. Key predictors were age, sex, and education level. This study developed ML-based risk assessment models that could become valuable clinical tools for evaluating lung cancer risk especially in never smokers.

4.1.2. *Predicting smoking-related postmenopausal osteoporosis*

Osteoporosis, a skeletal disorder marked by reduced bone strength and density, is particularly prevalent in post-menopausal women, with over 50 % developing post-menopausal osteoporosis (PMOP) due to hormonal changes [63]. However, the condition occurs twice as often in women who smoke compared to non-smokers, resulting in smoking-related postmenopausal osteoporosis (SRPO) [64]. High-throughput microarray technologies have revolutionized disease research by pinpointing genomic variations, complemented by weighted gene co-expression network analysis (WGCNA) to link gene modules with disease traits [65]. Applying these advancements to SRPO research holds significant promise, yet traditional methods struggle with the complexity of such large datasets. ML could bridge this gap, enhancing SRPO research by efficiently handling complex, high-dimensional data [66]. ML algorithms integrate genomic and transcriptomic data to pinpoint specific genetic variations and biomarkers [67]. Techniques like deep learning and ensemble methods improve predictive accuracy by uncovering intricate relationships among variables. Furthermore, ML has the potential to reveal new therapeutic targets by analyzing gene expression data to identify crucial pathways and networks relevant to SRPO [68]. Continuous adaptation and learning from new data enable ML models to provide ongoing insights, driving forward our understanding and treatment of SRPO. Using ML methods SVM-RFE and RF feature selection, Li et al. explored biomarkers and molecular pathways in SRPO, pinpointing six genes (HNRNPC, PFDN2, PSMC5, RPS16, TCEB2, UBE2V2) as genetic indicators [69]. Their findings not only identified potential biomarkers but also offered new insights into the molecular mechanisms driving SRPO. These findings offer specific genetic biomarkers (HNRNPC, PFDN2, PSMC5, RPS16, TCEB2, UBE2V2) for SRPO, which can advance traditional research by guiding further studies into disease mechanisms and facilitating targeted therapies. Clinically, these biomarkers could enhance early detection and personalized treatment strategies, improving outcomes for SRPO patients and potentially leading to the development of new diagnostic tools and therapies.

ML techniques have been effectively used to identify biomarkers and genetic profiles associated with smoking-related diseases. For instance, XGBoost has been employed to predict the likelihood of SiNCDs and identify specific gene isoforms linked to smoking. There is a need for more comprehensive studies that incorporate diverse populations to validate the predictive models further and assess their applicability in real-world settings. Future work could explore deep learning techniques to analyze complex genomic data, potentially leading to the discovery of novel biomarkers for personalized interventions. Additionally, integrating multi-omics data (genomics, proteomics, etc.) using ensemble methods could enhance the predictive accuracy of health outcomes.

5. Understanding passive smoke exposure

Secondhand and thirdhand smoke exposure can be classified as inactive or involuntary smoking, which pose significant public health risks, linked to respiratory infections, cardiovascular diseases, and cancer [70]. Secondhand smoke (SHS) is smoke that non-smokers involuntarily inhale from nearby smokers, whereas thirdhand smoke (THS) comprises residual contaminants that linger on surfaces and dust. Children are particularly vulnerable to exposure to both SHS and THS, posing significant health risks [71].

Traditional methods like observational studies and self-reported surveys, despite extensive research, suffer from accuracy issues, recall bias, high costs, and inefficiency with large datasets. ML offers promise by enhancing prediction of infant tobacco smoke exposure through advanced data analysis, improving understanding of less apparent sources like THS and less apparent routes like oral or dermal routes. Parks et al. highlighted ongoing challenges in reducing smoke exposure and the need for further exploration of SHS/THS impacts [71]. The study evaluated how well questionnaires could predict changes in urinary cotinine and 3HC levels. It found that the pervasiveness and persistence of the SHS and THS is so profound that modeling also could not predict more than half of the variation in urinary cotinine and 3HC levels in infants [71]. Despite low maternal smoking rates, high levels of cotinine (76 %) and 3HC (89 %) were detected in infants' urine, with questionnaire models explaining up to 41 % of variance [71]. Identified cut-points suggest significant SHS exposure in 23.5 % of infants [71]. Also, while THS is a recognized public health risk, traditional methods lack reliable biomarkers to distinguish it from SHS, limiting accurate assessments. A study by Merianos et al. aimed to improve differentiation between SHS and THS exposure in children using ML [72]. The model achieved prediction accuracies of 100 % for no/minimal exposure, 88 % for predominant THS exposure, and 71 % for mixed SHS and THS exposure [72]. Key predictors included the number of household smokers, serum cotinine, serum hydroxycotinine, and urinary 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol [72]. This demonstrates that ML can effectively distinguish between SHS and THS exposures, enhancing public health strategies for protecting children.

Smoking during pregnancy significantly impacts infant health, increasing risks of NCDs through DNA methylation, with studies identifying 185 altered CpG sites linked to maternal smoking, highlighting potential targets for understanding and mitigating these effects [73]. Rauschert et al. used ML to develop a DNA methylation score to identify fetal exposure to maternal smoking, using adolescent DNA data [74]. Tested in multiple cohorts, the score showed high sensitivity and specificity, outperforming existing non-ML based scores, and serves as a potential biomarker for assessing long-term health impacts of maternal smoking.

In addition to direct physiological effects, SHS has psychological impacts on non-smokers. Traditional studies struggle to identify key factors behind these psychological effects. Kim et al. used ML, specifically light gradient boosting machine (LGBM), to pinpoint stress perception, health status, and quality of life as key predictors of SHS-induced depression [75]. These findings highlight potential targets for depression-preventive interventions during public health crises.

ML has shown promise in distinguishing between secondhand and thirdhand smoke exposures, with studies achieving high prediction accuracies using various algorithms, such as LGBM. The current research does not extensively cover the long-term health impacts of THS and the psychological effects of SHS on non-smokers. Future research could leverage ML to handle and analyze large volumes of data from environmental sensors, air quality monitors, and wearable devices, providing accurate, continuous measurements of SHS and THS exposure. This real-time monitoring would enhance our understanding of exposure patterns and identifies high-risk environments and populations. ML also have the ability to improves predictive capabilities, identifying important predictors, forecasting potential exposure scenarios based on historical data, environmental conditions, and behaviors, enabling proactive protection measures. Furthermore, ML could personalize interventions by analyzing individual exposure data and health outcomes, tailoring strategies to specific needs. For instance, personalized recommendations can reduce SHS and THS exposure in homes and communities.

6. Predicting and improving smoking cessation outcomes

Smoking cessation programs aim to reduce tobacco-related diseases,

but success rates vary due to factors like motivation, support systems, nicotine dependence, and psychological aspects. In the U.S., 70 % of smokers attempt to quit, but only 7 % remain abstinent for over a year [76]. Though health care provider and patient discussion about quitting tobacco smoking have significant positive effect on helping patient to quit smoke, due to poor success rates, many physicians deprioritize smoking cessation discussions, creating a negative feedback loop [76, 77]. Understanding smoking behavior is crucial for effective interventions. Data-driven, personalized approaches can improve outcomes by predicting individual success and optimizing cessation services and nicotine replacement therapies. ML enhances these efforts by analyzing large datasets to identify patterns and correlations and tracking real-time data from wearable devices for personalized feedback. Predictive analytics forecast success and offer timely interventions, while reinforcement learning continuously improves strategies. Integrating ML leads to more effective, personalized interventions, enhancing success rates and public health. In a recent study, Issabakhsh et al. utilized ML to identify key determinants of smoking cessation using data from the PATH survey, achieving a prediction accuracy of 72 % for cessation between waves 1 and 2, and 70 % for cessation between waves 2 and 3 [78]. ML techniques, including random forest, gradient boosting machines, and the SHapley Additive explanation method, were employed to select important variables and assess their impact. The analysis found that factors such as recent e-cigarette use, lower prior cigarette use, older age at smoking initiation, shorter smoking duration, poly tobacco use, and higher BMI were significant predictors of successful smoking cessation.

6.1. Identifying high risk population for tobacco uses

Smoking behavior has a significant genetic component, with heritability as high as 50 %. Predicting individuals' predisposition to smoking through genomic profiles could help prevent those at risk from starting. Using ML methods, primarily SVM and RF, Xu et al. built models to predict smoking behavior from single nucleotide polymorphisms (SNPs) data, achieving high accuracy with a model that combined logistic and LASSO regression for feature selection [79]. The studies by Lai et al. and Hu et al. complement the previous study by offering personalized, precision medicine for smoking cessation [80] and providing data-driven insights into key factors in smoking discussions between health providers and patients [77]. These approaches can potentially improve success rates for individuals attempting to quit smoking. ANN-based ML prediction model has been used to readily available patient data to provide reliable, individualized smoking cessation success rates, thereby enhancing shared decision-making between physicians and patients [80]. Using data from 4875 patients at a medical center in Northern Taiwan, the model predicted smoking cessation outcomes over six months [80]. Thus, knowing both genetic predisposition and the likelihood of cessation success can greatly enhance the efficacy of smoking cessation programs. Bayesian Additive Regression Trees (BART), a ML technique, to recognize health care resource usage, smoking intensity and duration and smoking-related conditions were key determinants promoting discussions on smoking and facilitating smoking cessation, whereas the "usual suspects", age, gender, race and ethnicity were less important, and gender, in particular, had little effect on the likelihood of provider-patient discussions about smoking [77].

6.2. Predicting different aspects of smoking behavior

In a smoking cessation-related study, ML has been harnessed through a CNN-LSTM-based deep learning architecture named DeepPuff, which quantifies Respiratory Smoke Exposure Metrics (RSEM) by detecting smoke inhalations via breathing and hand gesture sensors from the Personal Automatic Cigarette Tracker v2 (PACT 2.0) [81]. The model achieved high precision in identifying smoke inhalations and provided reliable metrics for smoke exposure, comparable to those obtained through video annotation. The outcome demonstrates that the ML model, DeepPuff, is a reliable tool for measuring smoke exposure under free-living conditions, significantly contributing to the detailed assessment of smoking behavior and its health effects. In another study by Le et al., ML techniques, specifically RF paired with Recursive Feature Elimination and XGBoost, were used to identify predictors of smoking initiation among adults using the PATH study [82]. Notably, BMI and dental and oral health status emerged as robust predictors of smoking initiation, alongside other established risk factors. This research highlights the effectiveness of ML in uncovering new predictors and enhancing our understanding of tobacco use behaviors. Evans et al. utilized ML to identify and analyze distinct smoker profiles from a national survey, revealing four key groups: high-risk alcohol drinkers without children; single individuals in social housing with poor health and mental health; younger singles who have tried e-cigarettes and have poor mental health; and older couples with poor health [83]. These profiles differed notably from those of ex-smokers, who generally had better affluence, employment, and were older. This segmentation helps tailor targeted interventions and policies to specific smoker demographics, enhancing the effectiveness of smoking cessation efforts. Thakur et al. developed a ML-based framework using data from a wrist-wearable IMU sensor to accurately detect smoking activity among daily activities in real-time, achieving up to 98.7 % accuracy [84]. This ML approach enables just-in-time interventions for smoking cessation, aiding healthcare professionals in monitoring and supporting patients' quit efforts.

ML holds significant promise in advancing tobacco research, particularly in understanding the distinct epigenomic markers associated with different forms of smoking. A recent study on a Middle Eastern population, including never-smokers, cigarette-only smokers, and waterpipe-only smokers, used DNA methylome-wide profiling to reveal predominantly unique epigenetic markers for waterpipe smokers, distinct from those of cigarette smokers [85]. Remarkably, ML algorithms could accurately infer smoking forms with about 90 % accuracy based on DNA methylation patterns. These markers exhibited dose–response relationships with smoking extent and were validated through additional samples and independent technologies. By identifying markers enriched in regulatory regions and addiction-related pathways, the study highlights ML's potential to uncover specific epigenetic changes linked to addiction, aiding in the development of targeted interventions and prevention strategies for tobacco-related illnesses.

As mentioned in earlier section that vaping mostly associated with several risk factors, some research indicates that certain traits may be linked to adult smokers who successfully quit using e-cigarettes [86]. A study involving 889 adult smokers in Ontario used a gradient boosting machine model to identify key predictors of success in vaping-assisted smoking cessation [86]. The model, which achieved high performance with an ROC curve of 0.865 and a classification accuracy of 0.831, found that positive vaping experiences, fewer prior failed quit attempts, younger age, frequent vaping, and vaping shortly after waking were top predictors of perceived success [86]. This type of ML model could assist healthcare providers in recommending vaping for smoking cessation when it is likely to be beneficial. However, due to the risks of nicotine dependence and other health issues, vaping should be strictly limited and carefully considered, especially for adolescents.

6.3. ML powered apps

The low success rate in tobacco cessation is concerning. Nicotine cravings are the primary driver of smoking urges, but factors such as location, time, the presence of other smokers, and specific activities also play a significant role in triggering these urges. Studies show higher cravings in the mornings, evenings, at home, bars, and during media consumption. Effective cessation requires timely, targeted interventions.

Smoking cessation apps use data from smartphones and smartwatches to track these factors, providing just-in-time adaptive interventions (JITAI) through SMS, calls, or notifications, offering a cost-effective alternative to traditional methods [87]. Here, ML can enhance smoking cessation efforts by using real-time, environmentally contextual data to tailor interventions through Ecological Momentary Assessment (EMA) [88, 89]. Hebert et al. recently unrevealed strong predictors of first lapse during quit attempts [90]. Through the elastic net algorithm, the study presented five strongest predictors (i.e., perceived odds of smoking today, motivation to avoid smoking, confidence in ability to avoid smoking, cigarette availability, and urge to smoke) of the first lapse during quit attempt [90]. Similarly, Koslovsky et al. showed that Bayesian structural time series as a vigorous method to evaluate the risk factors connected with various stages of the tobacco smoking cessation process [88]. The model was effective in identifying 20 predictors (like, availability of cigarette, craving to smoke, location, behavioural and environmental factors etc.), their associations and their influence during smoking to non-smoking transition state or vice a versa. Dumortier et al. used Bayes, decision tree learning, and discriminant analysis on data from 349 smokers to accurately classify high craving states during quit attempts, highlighting ML's potential to aid smoking cessation and forecast urges via mobile health apps [91]. Another related recent study by Abo-Tabik et al. developed a deep learning-powered app that learns a smoker's routine and predicts smoking events. This app adjusts for individual differences among smokers and uses motion and geolocation data collected from mobile devices to forecast smoking occurrences [92]. However, it can be conclusively said that, combining passive automated data collection with ML models can enhance smoking cessation apps by providing just-in-time adaptive interventions during high-risk moments, thereby reducing lapses and preventing relapse, while researchers should prioritize high-quality sensor data over unreliable self-reported data.

In a recent study, Huang et al. harnessed ML to identify predictors of success in a depression-specific smoking cessation intervention using the Goal2Quit app [93]. By employing the LASSO to analyze baseline variables, the study determined factors such as time spent using the app and educational attainment [93]. Results showed that significant app usage and having an educational degree beyond high school were predictive of successful smoking reduction and depression improvement. This approach allowed for the personalization of digital cessation interventions, highlighting the importance of tailoring treatments based on user engagement and educational level.

A recent interesting study how ML can be utilized to play a critical role in this smoking cessation-related study by using supervised ML (SML) algorithms to analyze user data from a smoking cessation app [94]. The study aimed to identify app features that promote successful smoking cessation. By recording participants' app activity, demographic data, and tobacco use behaviors, the SML model was trained to predict the likelihood of cessation based on the use of different app features. The SML model demonstrated reasonable accuracy in predicting cessation, highlighting that patterns of app feature use could account for variance in smoking cessation beyond known predictors. This approach shows the potential of ML to optimize and enhance the effectiveness of smoking cessation apps.

Vera et al. used ML to identify key predictors of engagement with the Stop-Tabac smoking cessation app [95]. Algorithms helped reveal that the top predictors included the user's intention to stop smoking, dependence level, perceived helpfulness of the app, quitting success after one month, app usage after one month, group assignment (experimental vs. control), age, and years of smoking [95]. These insights can help tailor the app to user needs, improve enrolment, and enhance content. Etter et al. used ML algorithms to analyze data from 5293 daily smokers using the Stop-Tabac app, identifying key predictors of smoking cessation, reduction, and relapse at six months [96]. ML helped identify that tobacco dependence, motivation to quit, frequency of app use, perceived app usefulness, and nicotine medication use were significant predictors of smoking outcomes. This research provides valuable insights for improving smoking cessation apps and guiding future studies.

ML has been utilized to enhance smoking cessation outcomes by analyzing large datasets to identify key predictors of success. For example, Issabakhsh et al. achieved significant prediction accuracy using Random Forest and GBM. There is a lack of integration of behavioral and psychological factors into ML models, which could provide a more holistic understanding of smoking cessation. Another study by Perski et al. highlighting the need for combining unprompted and prompted data for effective intervention development [97]. This study used app-based data to develop ML models for predicting smoking lapses. Group-level models performed well (AUC = 0.969) but showed variability with new users [97]. Individual and hybrid models improved accuracy but were limited by data availability. Future studies could implement hybrid models combining ML with psychological theories to better predict cessation success. Additionally, exploring reinforcement learning for adaptive intervention strategies could lead to more effective cessation programs tailored to individual needs.

7. Effective intervention in the evolving smoking tobacco product landscape

Nicotine poses significant health risks, not only through traditional combustible products like cigarettes, cigars, and pipes, but also through evolving new age tobacco products like electronic nicotine delivery systems (ENDS), commonly known as vapes [98]. Experts considering the phenomenon as vaping epidemic [99]. Marketed under various names like e-cigarettes, vape pens, and pods by brands such as JUUL and Puff Bar, these devices heat nicotine-containing e-liquids to produce an aerosol for inhalation, mimicking smoking [100]. Despite their initial intent for smoking cessation, they've become popular among youth, with over 2.1 million U.S. teens currently using them [101]. This trend is alarming due to nicotine's adverse effects on attention, memory, and learning, potentially leading to addiction and subsequent tobacco use. Moreover, vapes contain harmful additive substances, toxic metals and chemicals similar to those in combustible cigarettes, contributing to irreversible lung damage [102]. The 2019 outbreak of vaping-associated pulmonary injury (EVALI) further underscores these risks, with numerous cases reported globally [103]. New-age tobacco products can be effectively managed using several ML approaches. ML can identify predictors for adolescent ENDS addiction, enhancing the precision of prevention and intervention strategies. Text-based sentiment analysis from social media can gauge public perceptions and concerns about these products, guiding targeted public health campaigns. ML can analyse online advertisements to detect and regulate misleading or youth-targeted promotions, thereby strengthening advertising regulations. Additionally, ML can detect toxic substances in ENDS, aiding regulatory efforts to ensure product safety and mitigate health risks associated with their use. These approaches are essential because traditional methods struggle with the vast volume of data involved, limiting their ability to derive timely and comprehensive insights. ML's capability to process and analyse big data enables efficient evaluation of predictors, sentiment analysis, advertisement content, and toxicity detection, significantly enhancing tobacco control efforts.

7.1. Identifying toxic substances in ENDS

Recent studies also highlight ML's promise in identifying the toxic effects of vaping when combined with traditional approaches. Vaping heats e-liquids to high temperatures, potentially creating harmful decomposition products. Kishimoto et al. uses a graph-convolutional neural network model predicted pyrolysis reactivity for 180 e-liquid flavors, generating 7307 products, which were refined using mass spectrometry data to identify 1169 molecular weight matches, revealing numerous toxic and hazardous compounds, thus aiding in understanding vaping's long-term health risks [104]. ML, particularly the CatBoost

algorithm, enhances detection of nicotine-containing e-liquids via surface-enhanced Raman scattering (SERS) [102]. This approach improves sensitivity and accuracy, enabling rapid on-site screening with portable devices and complementing central lab analyses. It holds potential for identifying prohibited additives, advancing tobacco control efforts.

7.2. Factors influencing youth smoking initiation

Intervention strategies aimed at preventing teenage ENDS use should target specific predictors unique to these products, such as digital media engagement and technology interest. Han et al. employed supervised ML, particularly penalized logistic regression, to identify these predictors from the Population Assessment of Tobacco and Health Study (PATH) [19]. Their findings highlight that frequent social media use significantly predicts ENDS use, distinct from other substance use behaviors, suggesting the influence of ENDS marketing on social platforms. Moreover, social media use predicts future cigarette smoking and experimentation with substances like alcohol and marijuana, underscoring broader risks associated with ENDS use. Recent studies further reinforce the impact of social media and identify additional predictors among tobacco-naive young adults, including susceptibility to ENDS, physical exercise, marijuana use, and susceptibility to cigarette use [105]. Shi et al.'s use of random forest algorithms supports this substance use predictors and reveals new insights such as the influence of school absences and significant ethnic interactions in predicting ever-vaping [106]. Vazquez et al. utilized the elastic net algorithm to validate individual and socioecological classifiers previously identified, such as substance use behaviors, perceptions of e-cigarette availability and risk, school-related factors like suspensions, and social influences like friends' behaviors [107].

Singh et al. used ML to identify predictors of vaping dependence over three months among daily and non-daily adolescent vapers. Key predictors for daily vapers included purchase location, pod usage duration, and nicotine vaping frequency; for non-daily vapers, predictors included race, sexual orientation, and heart disease treatment [108]. Fu et al. employed random forest to predict frequent vaping among adolescents, highlighting predictors such as higher past-month nicotine concentration in vape, frequent daily vaping sessions, and greater nicotine dependence [109]. These studies emphasize the importance of socioecological factors such as age, perceived discrimination, and race/-ethnicity in shaping adolescent attitudes towards vaping. Le also highlighted the influence of peer behaviors, household tobacco use, curiosity about ENDS, and perceptions of product safety [110].

These studies clearly indicate that social media usage and socioecological factors such as school-related issues, peer influence, ethnicity, and family attitudes toward tobacco play critical roles as predictors of adolescent ENDS use. These insights underscore ML's role in informing precise prevention strategies amid increasing ENDS use among youths, emphasizing the impact of family and social environments on adolescent tobacco behaviors. They reveal previously unknown insights and demonstrate ML's potential to enhance ENDS prevention efforts by swiftly analyzing complex socioecological variables, aiding in early intervention for at-risk youth in the growing e-cigarette epidemic. These ML insights can significantly contribute to targeted and effective tobacco control policies.

7.3. Predicting tobacco trends from social media data

7.3.1. Sentiment analysis from social media

Text-based sentiment analysis from social media helps prevent the vaping epidemic by monitoring public opinions and detecting trends in e-cigarette use. But traditional studies on public attitudes toward vaping are hindered by manual data identification. ML algorithms are able to handle this big data and can extract meaningful insights by analyze posts and comments. It can identify shifts in sentiment and misinformation, allowing authorities to respond promptly and tailor interventions. This real-time data shapes effective tobacco control policies and discourages vaping, especially among youth.

Hassan et al. used text mining of tweets during the EVALI outbreak to analyze public sentiment, demonstrating the effectiveness of ML in extracting insights relevant to tobacco control policy [111]. These methods can inform regulatory policies for ENDS. A study by Ren et al. used a stacking ensemble model that accurately identified vaping-related tweets with an F1-score of 0.97 [112]. They showed that how ML automation enhances the efficiency and insightfulness of analyzing social media data for public opinion and health surveillance in respect to new age tobacco product.

ML also have the ability to popularize the anti-vape opinion on social platform as shown by a very recent study [113]. Xie et al. demonstrates how ML, particularly deep-learning models and statistical techniques, enhances tobacco research by analyzing anti-vaping Instagram image posts to identify features associated with high user engagement [113]. By employing advanced ML algorithms, the study identified key image features and textual content that significantly correlate with increased likes and comments on anti-vaping posts. These findings highlight ML's capability to extract nuanced information from large datasets, offering insights into effective communication strategies for public health messaging on e-cigarette risks. By addressing the prevalence of pro-vaping content on social media, this research underscores the importance of leveraging ML to shape targeted interventions aimed at curbing youth vaping and promoting public health awareness.

7.3.2. Surveillance on online advertisement across popular sites

Surveillance of online advertisements helps prevent the vaping epidemic by identifying and removing youth-targeted marketing, ensuring compliance with advertising laws, and reducing exposure to pro-vaping messages. Increased scrutiny of youth vaping highlights social media's role, with unsupervised ML analysis of 70,725 Instagram posts revealing 3331 ENDS sales, primarily by individuals and retailers, often without age verification, underscoring the need for stricter enforcement [114]. ML, particularly DL classifiers like LSTM-CNN, shows superior performance in identifying and characterizing vaping-related tweets without requiring extensive annotations [115]. This capability supports the development of an effective X-based vaping surveillance system, overcoming the limitations of traditional classifiers. E-cigarette promotion on social media, especially JUUL pod vaporizers, surged alongside youth e-cigarette use. Kostygina et al. analyzed JUUL-related messages on Instagram using ML algorithms, finding that most commercial posts used recruitment and addiction appeals, rather than cessation-related messaging [116]. The study by Kong et al. effectively utilizes ML to analyze e-cigarette content on YouTube, identifying prevalent themes like product reviews and instructional videos, prominent e-cigarette products, and various marketing strategies such as discounts and sales promotions [117]. Murthy et al. illustrates how ML, specifically computer vision techniques like YOLOv7, contributes to tobacco research by effectively detecting e-cigarette-related content in visual media on TikTok [118]. By automating the detection of vaping devices, hands, and vapor clouds from a dataset of TikTok images, the model achieved high accuracy and recall rates, demonstrating its robustness in identifying e-cigarette imagery. Another very recent study showcases the efficacy of deep learning-based object detection in monitoring e-cigarette product presence across Instagram and TikTok. Using a DyHead model with a Swin-Large backbone, researchers accurately identified various e-cigarette-related objects in images and videos, revealing increasing trends in promotional content over time [119]. This automated approach offers scalable surveillance capabilities, crucial for informing tobacco regulatory science and aiding social media platforms in promptly moderating tobacco-related imagery to mitigate adolescent exposure online.

A recent study by Lakatos et al. even harnesses ML to uncover hidden threats regarding covert advertisements of tobacco [120]. This study

shows promise in detecting covert tobacco advertisements by enabling unbiased and reproducible analysis of tobacco-related media even increasing scope to strengthening the regulatory measures. An integrated model combining text and image processing, generative techniques, and human reinforcement achieves detection accuracies of 74 % for images and 98 % for text [120]. This approach leverages pre-trained multimodal deep learning models to identify smoking content across media formats, even with limited training data, while allowing expert intervention for enhanced accuracy.

These findings highlight how ML-driven computer vision systems can improve surveillance and regulatory efforts by quickly identifying and analyzing e-cigarette content on popular social media platforms. ML models accurately categorize video content and detect sales-related themes, which is vital for monitoring trends and understanding youth exposure to vaping imagery [120]. These information helps in developing targeted public health interventions to reduce the impact of e-cigarette use among young people. It also underscores the urgency of implementing stricter regulations to counter persuasive e-cigarette marketing aimed at youth online.

Based on current trends and research, it is reasonable to predict that ML will significantly address ENDS-related public health issues. By analyzing data from social media, sales platforms, and health records, ML can monitor trends, detect non-compliant products, assess health risks like EVALI, and provide personalized smoking cessation interventions. This will likely enhance regulatory oversight and inform effective decisions on product formulations, marketing claims, and health impacts, thereby improving public health responses to ENDS.

8. Economic impact of tobacco smoking

Nargis et al. estimated state-level economic losses attributable to cigarette smoking in the USA using a dynamic macroeconomic model of personal income per capita, analyzing data from 2011 to 2020 with a mixed-effects, generalized linear, dynamic panel data model [121]. The study found substantial income losses, with a national annual combined loss of $436.7 billion and a cumulative loss of $864.5 billion in 2020 due to smoking [121]. While this study did not use ML, ML could enhance such economic models by analyzing large datasets more effectively, uncovering hidden patterns, refining estimates, and enabling real-time adjustments as new data becomes available. Smoking leads to significant economic losses in the USA. Equitable tobacco control measures can greatly enhance macroeconomic performance both short and long term by lowering health costs and preventing productivity declines.

9. Conclusion

ML has revolutionized tobacco research by providing innovative solutions to the complexities of tobacco use and its associated health impacts. With its ability to analyze extensive datasets and apply advanced algorithms, ML uncovers patterns, predicts outcomes, and personalizes interventions. This capability enhances our understanding of tobacco-related health risks and improves cessation strategies. Significant insights driven by ML include the identification of genetic predispositions to smoking-related diseases, predictions of smoking behaviors, and assessments of secondhand and thirdhand smoke exposure. These findings not only deepen our comprehension of the biological mechanisms underlying tobacco addiction but also aid in developing targeted and personalized interventions that can lead to improved health outcomes. Despite these advancements, ML in tobacco research faces challenges. There is a lack of personalized health impact predictions for SiNCDs beyond lung cancer and SRPO, despite the existence of numerous other SiNCDs that should also be considered for prediction using ML models in the smoking population. Also on the technical side, the effectiveness of ML models depends on high-quality data; inaccurate or biased datasets can lead to misleading results and ineffective interventions. Additionally, many ML models, especially deep learning algorithms, are difficult to interpret, which complicates understanding their decision-making processes. Ethical issues regarding privacy, consent, and data misuse also arise. To address these challenges, researchers should focus on improving data quality, fostering institutional collaborations for better data sharing, enhancing model transparency through interpretable techniques, and establishing clear ethical guidelines for data use and privacy. Despite these drawbacks, the future of ML in tobacco research appears promising. The integration of wearable devices and environmental sensors can provide continuous data on tobacco exposure, enabling real-time monitoring and personalized interventions. ML algorithms will increasingly support the development of tailored cessation programs that adapt to individual behaviors and health profiles. Additionally, ML can assist in monitoring and regulating emerging tobacco products, such as e-cigarettes, by analyzing social media trends and sales data to identify potential public health risks. Future research will benefit from cross-disciplinary collaboration among data scientists, public health experts, and clinicians, fostering innovation and developing comprehensive strategies to combat tobacco use. In conclusion, the integration of ML into tobacco research signifies a paradigm shift, offering advanced tools to understand and address tobacco use. Embracing ML while addressing its challenges can enhance public health outcomes and advance the global fight against tobacco use in a responsible and ethical manner.

Future investigations in tobacco research should focus on adopting cutting-edge ML techniques to uncover untapped opportunities and enhance current methodologies. For predictive tasks, transformer-based models like Vision Transformers (ViTs) and Temporal Fusion Transformers (TFTs) can process diverse datasets to predict disease risks and understand patterns of health decline over time [122,123]. Behavioral analyses could utilize unsupervised methods, such as Variational Autoencoders (VAEs), to segment smoker profiles, while reinforcement learning strategies like deep Q-networks (DQN) and proximal policy optimization (PPO) could help predict and influence behavioral changes in response to interventions [124,125]. Moreover, advanced computer vision algorithms, including Vision Transformers, can significantly improve the detection of adulterants and counterfeit tobacco products. Genomic research can benefit from Graph Neural Networks (GNNs) to examine intricate gene-environment interactions and facilitate the development of nicotine addiction treatments. Public health studies can employ sophisticated natural language processing models, such as GPT-4, for monitoring tobacco-related discussions on digital platforms and utilize causal ML methods like Structural Equation Models (SEMs) to assess the outcomes of policy interventions [126,127]. By incorporating Explainable AI and fairness-focused frameworks, these ML applications can promote ethical, transparent, and impactful advancements in tobacco research.

Funding

No funding was received for the study.

CRediT authorship contribution statement

Nabanita Ghosh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Conceptualization. **Krishnendu Sinha:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Investigation, Conceptualization. **Parames Sil:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used freely available LLMs in order to improve language and readability, with caution.

After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We extend our gratitude to those who have never smoked for their commitment to health, family and society, and to those who have quit smoking for their dedication to personal and public well-being. We also appreciate the efforts of individuals, groups, and organizations working towards a tobacco-free environment.

Data Availability

Data will be made available on request.

References

[1] X. Gui, Z. Yang, M.D. Li, Effect of cigarette smoke on gut microbiota: state of knowledge, Front Physiol. 12 (2021) 673341, https://doi.org/10.3389/FPHYS.2021.673341/BIBTEX.
[2] T. Behl, I. Kaur, A. Sehgal, P.S. Khandige, M. Imran, M. Gulati, M. Khalid Anwer, G.M. Elossaily, N. Ali, P. Wal, A. Gasmi, The link between Alzheimer's disease and stroke: a detrimental synergism, Ageing Res Rev. 99 (2024) 102388, https://doi.org/10.1016/J.ARR.2024.102388.
[3] R.J. Bonnie, K. Stratton, L.Y. Kwan, C. on the P.H.I. of R. the M.A. for P.T. Products, B. on P.H. and P.H. Practice, I. of Medicine, Eff. Tob. Use Health (2015). ⟨https://www.ncbi.nlm.nih.gov/books/NBK310413/⟩. accessed July 22, 2024.
[4] F. Yangui, A. Touil, S. Antit, L. Zakhama, M.R. Charfi, COPD prevalence in smokers with stable ischemic heart disease: a cross-sectional study in Tunisia, Respir. Med 179 (2021) 106335, https://doi.org/10.1016/j.rmed.2021.106335.
[5] Tobacco, (n.d.). ⟨https://www.who.int/news-room/fact-sheets/detail/tobacco⟩ (accessed August 12, 2022).
[6] X. Dai, E. Gakidou, A.D. Lopez, Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action, Tob. Control 31 (2022) 129–137, https://doi.org/10.1136/TOBACCOCONTROL-2021-056535.
[7] Tobacco control, (n.d.). ⟨https://www.who.int/data/gho/data/themes/theme-details/GHO/tobacco-control⟩ (accessed August 12, 2022).
[8] N.A. Rigotti, G.R. Kruse, J. Livingstone-Banks, J. Hartmann-Boyce, Treatment of tobacco smoking: a review, JAMA 327 (2022) 566–577, https://doi.org/10.1001/JAMA.2022.0395.
[9] C. Zhu, S. Young-Soo, R. Beaglehole, Tobacco control in China: small steps towards a giant leap, Lancet 379 (2012) 779–780, https://doi.org/10.1016/S0140-6736(11)61933-8.
[10] N. Shah, M. Nali, C. Bardier, J. Li, J. Maroulis, R. Cuomo, T.K. Mackey, Applying topic modelling and qualitative content analysis to identify and characterise ENDS product promotion and sales on Instagram, Tob. Control (2021) tobaccocontrol-2021-056937, https://doi.org/10.1136/TOBACCOCONTROL-2021-056937.
[11] E.E. Litsa, P. Das, L.E. Kavraki, Machine learning models in the prediction of drug metabolism: challenges and future perspectives, Expert Opin. Drug Metab. Toxicol. 17 (2021) 1245–1247, https://doi.org/10.1080/17425255.2021.1998454.
[12] A.S. Hatoum, F.R. Wendt, M. Galimberti, R. Polimanti, B. Neale, H.R. Kranzler, J. Gelernter, H.J. Edenberg, A. Agrawal, Ancestry may confound genetic machine learning: candidate-gene prediction of opioid use disorder as an example, Drug Alcohol Depend. 229 (2021) 109115, https://doi.org/10.1016/J.DRUGALCDEP.2021.109115.
[13] Y. Ren, D. Wu, A. Singh, E. Kasson, M. Huang, P. Cavazos-Rehg, Automated detection of vaping-related tweets on twitter during the 2019 EVALI outbreak using machine learning classification, Front Big Data 5 (2022) 5, https://doi.org/10.3389/FDATA.2022.770585/BIBTEX.
[14] K. Sinha, N. Ghosh, A review on the recent advancements in machine learning-assisted tobacco research, NIPES - J. Sci. Technol. Res. 6 (2024) 2024–2055, https://doi.org/10.5281/ZENODO.11223324.
[15] A. Géron, Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems, O'Reilly Media (2019) 851. ⟨https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/⟩ (accessed June 15, 2022).
[16] A.L. Samuel, Some studies in machine learning using the game of checkers, IBM J. Res Dev. 3 (1959) 210–229, https://doi.org/10.1147/RD.33.0210.
[17] Master Machine Learning Algorithms, (n.d.). ⟨https://machinelearningmastery.com/master-machine-learning-algorithms/⟩ (accessed August 21, 2022).
[18] Practical Machine Learning Problems, (n.d.). ⟨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⟩ (accessed August 21, 2022).
[19] D.H. Han, S.H. Lee, S. Lee, 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 (Balt.) 145 (2021), https://doi.org/10.1016/J.YPMED.2021.106418.
[20] S. Lee, D.H. Han, A. Chow, D.C. Seo, A prospective longitudinal relation between elevated use of electronic devices and use of electronic nicotine delivery systems, Addict. Behav. 98 (2019) 106063, https://doi.org/10.1016/J.ADDBEH.2019.106063.
[21] A. Burkov, Machine learning engineering, 2020. ⟨https://www.saint-gobain.co.in/sites/saint-gobain.co.in/files/webform/apply_for_indec/_sid_/machine-learning-engineering-andriy-burkov-pdf-download-free-book-c6498c8.pdf⟩ (accessed June 15, 2022).
[22] S. Badillo, B. Banfai, F. Birzele, I.I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, J.D. Zhang, An introduction to machine learning, Clin. Pharm. Ther. 107 (2020) 871–885, https://doi.org/10.1002/CPT.1796.
[23] R. Fu, A. Kundu, N. Mitsakakis, T. Elton-Marshall, W. Wang, S. Hill, S.J. Bondy, H. Hamilton, P. Selby, R. Schwartz, M.O. Chaiton, Machine learning applications in tobacco research: a scoping review, Tob. Control (2021) tobaccocontrol-2020-056438, https://doi.org/10.1136/TOBACCOCONTROL-2020-056438.
[24] P. Kaviani, M.S. Dhotre, Short survey on naive bayes algorithm, Int. J. Adv. Eng. Res. Dev. 4 (2017).
[25] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, C. Fernandez-Lozano, A review on machine learning approaches and trends in drug discovery, Comput. Struct. Biotechnol. J. 19 (2021) 4538–4558, https://doi.org/10.1016/J.CSBJ.2021.08.011.
[26] I. Rish, I. Rish, Empir. Study naive bayes Classif. (2001). ⟨http://citeseerx.ist.psu.edu/viewdoc/summary?doi= 10.1.1.330.2788⟩ (accessed June 18, 2022).
[27] LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S, Philos. Trans. R. Soc. Lond. 53 (1763) 370–418, https://doi.org/10.1098/RSTL.1763.0053.
[28] Y. Huang, L. Li, Naive Bayes classification algorithm based on small sample set, CCIS2011 - Proceedings: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (2011) 34–39. https://doi.org/10.1109/CCIS.2011.6045027.
[29] S. Raschka, STAT 479: Machine Learning Lecture Notes, (2018). ⟨http://stat.wisc.edu/⟩~sraschka/teaching/stat479-fs2018/ (accessed June 17, 2022).
[30] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (1992) 175–185, https://doi.org/10.1080/00031305.1992.10475879.
[31] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27, https://doi.org/10.1109/TIT.1967.1053964.
[32] T. Hastie, R. Tibshirani, J. Friedman, Elem. Stat. Learn. (2009), https://doi.org/10.1007/978-0-387-84858-7.
[33] L. Breiman, Bagging predictors, 24, Mach. Learn. 1996 24 (2) (1996) 123–140, https://doi.org/10.1007/BF00058655.
[34] G. Louppe, P. Geurts, Ensembles on random patches (LNAI), Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.) 7523 (2012) 346–361, https://doi.org/10.1007/978-3-642-33460-3_28/COVER/.
[35] J.F.-A. of statistics, undefined 2001, Greedy function approximation: a gradient boosting machine, JSTOR 29 (2001) 1189–1232. ⟨https://www.jstor.org/stable/2699986⟩ (accessed June 23, 2022).
[36] J. Zhu, J. Zhu, H. Zou, S. Rosset, T. Hastie, Multi-class AdaBoost, Stat. Its Interface 2 (2009) 349–360. ⟨http://citeseerx.ist.psu.edu/viewdoc/summary?doi= 10.1.1.158.4221⟩ (accessed June 23, 2022).
[37] XGBoost Documentation — xgboost 1.6.1 documentation, (n.d.). ⟨https://xgboost.readthedocs.io/en/stable/⟩ (accessed August 21, 2022).
[38] K. Davagdorj, V.H. Pham, N. Theera-Umpon, K.H. Ryu, XGBoost-based framework for smoking-induced noncommunicable disease prediction, Int J. Environ. Res Public Health 17 (2020) 1–22, https://doi.org/10.3390/IJERPH17186513.
[39] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, 2000. ⟨https://books.google.com/books?hl=en&lr= &id= _PXJn_cxv0AC&oi=fnd&pg=PR9&ots=xTRi5IYt0h&sig=IkBAWd3LfgcSbrHt4SFaZaZR8E4⟩ (accessed June 16, 2022).
[40] C. Campbell, Y. Ying, Learning with support vector machines, Synth. Lect. Artif. Intell. Mach. Learn. 10 (2011) 1–95, https://doi.org/10.2200/S00324ED1V01Y201102AIM010.
[41] Y. Ren, D. Wu, A. Singh, E. Kasson, M. Huang, P. Cavazos-Rehg, Automated detection of vaping-related tweets on twitter during the 2019 EVALI outbreak using machine learning classification, Front Big Data 5 (2022) 5, https://doi.org/10.3389/FDATA.2022.770585/BIBTEX.
[42] P. Rzehak, R. Saffery, E. Reischl, M. Covic, S. Wahl, V. Grote, A. Xhonneux, J. P. Langhendries, N. Ferre, R. Closa-Monasterolo, E. Riva, P. Socha, D. Gruszfeld, B. Koletzko, P. Goyens, C. Carlier, J. Hoyos, P. Poncelet, E. Dain, J.N. Van Hees, F. Martin, J. Escribano, V. Luque, G. Mendez, M. Zaragoza-Jordana, M. Giovannini, C. Agostoni, S. Scaglioni, E. Verduci, F. Vecchi, A.Re Dionigi, J. Socha, A. Stolarczyk, A. Dobrzanska, R. Janas, E. Perrin, R. Von Kries, H. Groebe, A. Reith, R. Hofmann, M. Weber, S. Schiess, J. Beyer, M. Fritsch,

U. Handel, I. Pawellek, S. Verwied-Jorky, I. Hannibal, H. Demmelmair, G. Haile, M. Theurich, 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 (2016) e0155554, https://doi.org/10.1371/JOURNAL.PONE.0155554.
[43] L.S. Moulin, A.P. Alves Da Silva, M.A. El-Sharkawi, R.J. Marks Ii, Support vector machines for transient stability analysis of large-scale power systems, IEEE Trans. POWER Syst. 19 (2004), https://doi.org/10.1109/TPWRS.2004.826018.
[44] F. Chollet, Deep Learn. Python Second Ed. Deep Learn. Python (2021).
[45] Deep Neural Networks | Kaggle, (n.d.). ⟨https://www.kaggle.com/code/ryanholbrook/deep-neural-networks⟩ (accessed June 18, 2022).
[46] Y. Bengio, I. Goodfellow, A. Courville, Deep Learn (2017). ⟨https://www.academia.edu/download/62266271/Deep_Learning20200303-80130-1s42zvt.pdf⟩ (accessed June 18, 2022).
[47] C.M. Bishop, Pattern Recogniton and Machine Learning, Springer-Verlag, New York, 2006. ⟨https://www.springer.com/gp/book/9780387310732⟩ (accessed June 18, 2022).
[48] G. Cybenko, Approximation by superpositions of a sigmoidal function, 1989, Math. Control, Signals Syst. 2 (4) (1989) 303–314, https://doi.org/10.1007/BF02551274.
[49] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, 521, Nature 2015 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[50] R. Fu, A. Kundu, N. Mitsakakis, T. Elton-Marshall, W. Wang, S. Hill, S.J. Bondy, H. Hamilton, P. Selby, R. Schwartz, M.O. Chaiton, Machine learning applications in tobacco research: a scoping review, Tob. Control 32 (2023) 99–109, https://doi.org/10.1136/TOBACCOCONTROL-2020-056438.
[51] F. Huang, Q. Ma, J. Ren, J. Li, F. Wang, T. Huang, Y.D. Cai, Identification of smoking-associated transcriptome aberration in blood with machine learning methods, Biomed. Res Int 2023 (2023), https://doi.org/10.1155/2023/5333361.
[52] J.K. Yoon, S. Park, K.H. Lee, D. Jeong, J. Woo, J. Park, S.M. Yi, D. Han, C.G. Yoo, S. Kim, C.H. Lee, Machine learning-based proteomics reveals ferroptosis in COPD patient-derived airway epithelial cells upon smoking exposure, J. Korean Med. Sci. 38 (2023), https://doi.org/10.3346/JKMS.2023.38.E220.
[53] A. Budreviciute, S. Damiati, D.K. Sabir, K. Onder, P. Schuller-Goetzburg, G. Plakys, A. Katileviciute, S. Khoja, R. Kodzius, Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors, Front Public Health 8 (2020) 574111, https://doi.org/10.3389/FPUBH.2020.574111/PDF.
[54] B. Cao, F. Bray, A. Ilbawi, I. Soerjomataram, Effect on longevity of one-third reduction in premature mortality from non-communicable diseases by 2030: a global analysis of the Sustainable Development Goal health target, Lancet Glob. Health 6 (2018) e1288–e1296, https://doi.org/10.1016/S2214-109X(18)30411-X.
[55] M.A. Mallah, T. Soomro, M. Ali, S. Noreen, N. Khatoon, A. Kafle, F. Feng, W. Wang, M. Naveed, Q. Zhang, Cigarette smoking and air pollution exposure and their effects on cardiovascular diseases, Front Public Health 11 (2023) 967047, https://doi.org/10.3389/FPUBH.2023.967047/BIBTEX.
[56] W. Kopp, Pathogenesis of (smoking-related) non-communicable diseases—evidence for a common underlying pathophysiological pattern, Front Physiol. 13 (2022) 1037750, https://doi.org/10.3389/FPHYS.2022.1037750.
[57] R. Wang, Y. Qiang, X. Gao, Q. Yang, B. Li, Prevalence of non-communicable diseases and its association with tobacco smoking cessation intention among current smokers in Shanghai, China, Tob. Induc. Dis. 20 (2022), https://doi.org/10.18332/TID/155828.
[58] K. Davagdorj, V.H. Pham, N. Theera-Umpon, K.H. Ryu, XGBoost-based framework for smoking-induced noncommunicable disease prediction, Int J. Environ. Res Public Health 17 (2020) 1–22, https://doi.org/10.3390/IJERPH17186513.
[59] R.A. Pleasants, M.P. Rivera, S.L. Tilley, S.P. Bhatt, Both duration and pack-years of tobacco smoking should be used for clinical practice and research, Ann. Am. Thorac. Soc. 17 (2020) 804–806, https://doi.org/10.1513/ANNALSATS.202002-133VP/SUPPL_FILE/DISCLOSURES.PDF.
[60] R. Chen, J. Lin, Identification of feature risk pathways of smoking-induced lung cancer based on SVM, PLoS One 15 (2020) e0233445, https://doi.org/10.1371/JOURNAL.PONE.0233445.
[61] E. Nemlander, A. Rosenblad, E. Abedi, S. Ekman, J. Hasselström, L.E. Eriksson, A. C. Carlsson, Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers, PLoS One 17 (2022) e0276703, https://doi.org/10.1371/JOURNAL.PONE.0276703.
[62] S. Jain, V. Jain, Analytics on Risk Factors Correlated to Non-Communicable Diseases using Machine Learning, Proceedings of 2022 IEEE International Conference on Current Development in Engineering and Technology, CCET 2022 (2022). https://doi.org/10.1109/CCET56606.2022.10080674.
[63] Y. Geng, R. Shao, T. Xu, L. Zhang, Identification of a potential signature to predict the risk of postmenopausal osteoporosis, Gene 894 (2024) 147942, https://doi.org/10.1016/J.GENE.2023.147942.
[64] C. Oncken, S. Allen, M. Litt, A. Kenny, H. Lando, A. Allen, E. Dornelas, Exercise for smoking cessation in postmenopausal women: a randomized, controlled trial, Nicotine Tob. Res 22 (2020) 1587–1595, https://doi.org/10.1093/NTR/NTZ176.
[65] P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis, BMC Bioinforma. 9 (2008) 1–13, https://doi.org/10.1186/1471-2105-9-559/FIGURES/4.
[66] K. Huang, C. Xiao, L.M. Glass, C.W. Critchlow, G. Gibson, J. Sun, Machine learning applications for therapeutic tasks with genomics data, Patterns 2 (2021) 100328, https://doi.org/10.1016/J.PATTER.2021.100328.
[67] Y. You, X. Lai, Y. Pan, H. Zheng, J. Vera, S. Liu, S. Deng, L. Zhang, Artificial intelligence in cancer target identification and drug discovery, 7, Signal Transduct. Target. Ther. 2022 7 (1) (2022) 1–24, https://doi.org/10.1038/s41392-022-00994-0.
[68] H. Chen, F.J. King, B. Zhou, Y. Wang, C.J. Canedy, J. Hayashi, Y. Zhong, M. W. Chang, L. Pache, J.L. Wong, Y. Jia, J. Joslin, T. Jiang, C. Benner, S.K. Chanda, Y. Zhou, Drug target prediction through deep learning functional representation of gene signatures, 15, Nat. Commun. 2024 15 (1) (2024) 1–15, https://doi.org/10.1038/s41467-024-46089-y.
[69] S. Li, B. Chen, H. Chen, Z. Hua, Y. Shao, H. Yin, 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 (2021) e0257343, https://doi.org/10.1371/JOURNAL.PONE.0257343.
[70] G. Carreras, A. Lugo, S. Gallus, B. Cortini, E. Fernández, M.J. López, J.B. Soriano, A. López-Nicolás, S. Semple, G. Gorini, Y. Castellano, M. Fu, M. Ballbè, B. Amalia, O. Tigova, X. Continente, T. Arechavala, E. Henderson, X. Liu, C. Bosetti, E. Davoli, P. Colombo, R. O'Donnell, R. Dobson, L. Clancy, S. Keogan, H. Byrne, P. Behrakis, A. Tzortzi, C. Vardavas, V.K. Vyzikidou, G. Bakellas, G. Mattiampa, R. Boffi, A. Ruprecht, C. De Marco, A. Borgini, C. Veronese, M. Bertoldi, A. Tittarelli, S. Verdi, E. Chellini, M. Trapero-Bertran, D.C. Guerrero, C. Radu-Loghin, D. Nguyen, P. Starchenko, J. Ancochea, T. Alonso, M.T. Pastor, M. Erro, A. Roca, P. Pérez, Burden of disease attributable to second-hand smoke exposure: a systematic review, Prev. Med (Balt.) 129 (2019) 105833, https://doi.org/10.1016/J.YPMED.2019.105833.
[71] J. Parks, K.E. McLean, L. McCandless, R.J. de Souza, J.R. Brook, J. Scott, S. E. Turvey, P.J. Mandhane, A.B. Becker, M.B. Azad, T.J. Moraes, D.L. Lefebvre, M. R. Sears, P. Subbarao, T.K. Takaro, Assessing secondhand and thirdhand tobacco smoke exposure in Canadian infants using questionnaires, biomarkers, and machine learning, 32, J. Expo. Sci. Environ. Epidemiol. 2021 32 (1) (2021) 112–123, https://doi.org/10.1038/s41370-021-00350-4.
[72] A.L. Merianos, E.M. Mahabee-Gittens, T.M. Stone, R.A. Jandarov, L. Wang, D. Bhandari, B.C. Blount, G.E. Matt, Distinguishing exposure to secondhand and thirdhand tobacco smoke among U.S. children using machine learning: NHANES 2013-2016, Environ. Sci. Technol. 57 (2023) 2042–2053, https://doi.org/10.1021/ACS.EST.2C08121/ASSET/IMAGES/MEDIUM/ES2C08121_0005.GIF.
[73] C.A. Markunas, Z. Xu, S. Harlid, P.A. Wade, R.T. Lie, J.A. Taylor, A.J. Wilcox, Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy, Environ. Health Perspect. 122 (2014) 1147–1153, https://doi.org/10.1289/EHP.1307892.
[74] S. Rauschert, P.E. Melton, A. Heiskala, V. Karhunen, G. Burdge, J.M. Craig, K. M. Godfrey, K. Lillycrop, T.A. Mori, L.J. Beilin, W.H. Oddy, C. Pennell, M. R. Järvelin, S. Sebert, R.C. Huang, 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 (2020) 1–11, https://doi.org/10.1289/EHP6076.
[75] N.H. Kim, M. Kim, J.S. Han, H. Sohn, B. Oh, J.W. Lee, S. Ahn, Machine-learning model for predicting depression in second-hand smokers in cross-sectional data using the Korea National Health and Nutrition Examination Survey, Digit Health 10 (2024), https://doi.org/10.1177/20552076241257046.
[76] P. Caponnetto, R. Polosa, Common predictors of smoking cessation in clinical practice, Respir. Med. 102 (2008) 1182–1192, https://doi.org/10.1016/J.RMED.2008.02.017.
[77] L. Hu, L. Li, J. Ji, Machine learning to identify and understand key factors for provider-patient discussions about smoking, Prev. Med. Rep. 20 (2020) 101238, https://doi.org/10.1016/J.PMEDR.2020.101238.
[78] M. Issabakhsh, L.M. Sánchez-Romero, T.T.T. Le, A.C. Liber, J. Tan, Y. Li, R. Meza, D. Mendez, D.T. Levy, Machine learning application for predicting smoking cessation among US adults: an analysis of waves 1-3 of the PATH study, PLoS One 18 (2023), https://doi.org/10.1371/JOURNAL.PONE.0286883.
[79] Y. Xu, L. Cao, X. Zhao, Y. Yao, Q. Liu, B. Zhang, Y. Wang, Y. Mao, Y. Ma, J.Z. Ma, T.J. Payne, M.D. Li, L. Li, Prediction of smoking behavior from single nucleotide polymorphisms with machine learning approaches, Front Psychiatry 11 (2020), https://doi.org/10.3389/FPSYT.2020.00416/PDF.
[80] C.C. Lai, W.H. Huang, B.C.C. Chang, L.C. Hwang, Development of machine learning models for prediction of smoking cessation outcome, Int J. Environ. Res Public Health 18 (2021) 1–10, https://doi.org/10.3390/IJERPH18052584.
[81] P. Belsare, V.Y. Senyurek, M.H. Imtiaz, S.T. Tiffany, E. Sazonov, DeepPuff: utilizing deep learning for smoking behavior identification in free-living environment, Annu Int Conf. IEEE Eng. Med Biol. Soc. 2023 (2023), https://doi.org/10.1109/EMBC40787.2023.10340528.
[82] T.T.T. Le, M. Issabakhsh, Y. Li, L.M. Sánchez-Romero, J. Tan, R. Meza, D. Levy, D. Mendez, Are the relevant risk factors being adequately captured in empirical studies of smoking initiation? A machine learning analysis based on the population assessment of tobacco and health study, Nicotine Tob. Res. 25 (2023) 1481–1488, https://doi.org/10.1093/NTR/NTAD066.
[83] A. Evans, R. Hughes, L. Nolan, K. Little, L. Newbury-Davies, A.R. Davies, Profiles of tobacco smokers and ex-smokers in a large-scale random sample survey across Wales: an unsupervised machine-learning cluster analysis, Lancet 402 (1) (2023) S7, https://doi.org/10.1016/S0140-6736(23)02070-6.
[84] S.S. Thakur, P. Poddar, R.B. Roy, Real-time prediction of smoking activity using machine learning based multi-class classification model, Multimed. Tools Appl. 81 (2022) 14529–14551, https://doi.org/10.1007/S11042-022-12349-6/TABLES/9.
[85] Z. Awada, V. Cahais, C. Cuenin, R. Akika, A.L. Silva Almeida Vicente, M. Makki, H. Tamim, Z. Herceg, N. Khoueiry Zgheib, A. Ghantous, Waterpipe and cigarette

epigenome analysis reveals markers implicated in addiction and smoking type inference, Environ. Int 182 (2023) 108260, https://doi.org/10.1016/J.ENVINT.2023.108260.
[86] R. Fu, R. Schwartz, N. Mitsakakis, L.M. Diemert, S. O'Connor, J.E. Cohen, Predictors of perceived success in quitting smoking by vaping: a machine learning approach, PLoS One 17 (2022) e0262407, https://doi.org/10.1371/JOURNAL.PONE.0262407.
[87] M. Abo-Tabik, Y. Benn, N. Costen, Are machine learning methods the future for smoking cessation apps? Sens. (Basel) 21 (2021) https://doi.org/10.3390/S21134254.
[88] M.D. Koslovsky, M.D. Swartz, W. Chan, L. Leon-Novelo, A.V. Wilkinson, D. E. Kendzor, M.S. Businelle, Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation, Biometrics 74 (2018) 636–644, https://doi.org/10.1111/BIOM.12792.
[89] K.K. Mak, K. Lee, C. Park, Applications of machine learning in addiction studies: a systematic review, Psychiatry Res. 275 (2019) 53–60, https://doi.org/10.1016/J.PSYCHRES.2019.03.001.
[90] E.T. Hébert, R. Suchting, C.K. Ra, A.C. Alexander, D.E. Kendzor, D.J. Vidrine, M. S. Businelle, Predicting the first smoking lapse during a quit attempt: a machine learning approach, Drug Alcohol Depend. 218 (2021) 108340, https://doi.org/10.1016/J.DRUGALCDEP.2020.108340.
[91] A. Dumortier, E. Beckjord, S. Shiffman, E. Sejdić, Classifying smoking urges via machine learning, Comput. Methods Prog. Biomed. 137 (2016) 203–213, https://doi.org/10.1016/J.CMPB.2016.09.016.
[92] M. Abo-Tabik, N. Costen, J. Darby, Y. Benn, Towards a smart smoking cessation app: A 1D-CNN model predicting smoking events, Vol. 20, Page 1099, Sensors 2020 20 (2020) 1099, https://doi.org/10.3390/S20041099.
[93] S. Huang, A. Wahlquist, J. Dahne, Individual predictors of response to a behavioral activation-based digital smoking cessation intervention: a machine learning approach, Subst. Use Misuse (2024), https://doi.org/10.1080/10826084.2024.2369155.
[94] L.N. Siegel, K.P. Wiseman, A. Budenz, Y. Prutzman, Identifying patterns of smoking cessation app feature use that predict successful quitting: secondary analysis of experimental data leveraging machine learning, JMIR AI 3 (2024) e51756, https://doi.org/10.2196/51756.
[95] G. Vera Cruz, Y. Khazaal, J.F. Etter, Predicting the users' level of engagement with a smartphone application for smoking cessation: randomized trial and machine learning analysis, Eur. Addict. Res 29 (2023) 171–181, https://doi.org/10.1159/000530111.
[96] J.F. Etter, G. Vera Cruz, Y. Khazaal, Predicting smoking cessation, reduction and relapse six months after using the Stop-Tabac app for smartphones: a machine learning analysis, BMC Public Health 23 (2023), https://doi.org/10.1186/S12889-023-15859-6.
[97] O. Perski, K. Li, N. Pontikos, D. Simons, S.P. Goldstein, F. Naughton, J. Brown, Classification of lapses in smokers attempting to stop: a supervised machine learning approach using data from a popular smoking cessation smartphone app, Nicotine Tob. Res. 25 (2023) 1330, https://doi.org/10.1093/NTR/NTAD051.
[98] J. Drope, Z. Cahn, R. Kennedy, A.C. Liber, M. Stoklosa, R. Henson, C.E. Douglas, J. Drope, Key issues surrounding the health impacts of electronic nicotine delivery systems (ENDS) and other sources of nicotine, CA Cancer J. Clin. 67 (2017) 449–471, https://doi.org/10.3322/CAAC.21413.
[99] S.W. Huey, M.H. Granitto, Smoke screen: The teen vaping epidemic uncovers a new concerning addiction, J. Am. Assoc. Nurse Pr. 32 (2020) 293–298, https://doi.org/10.1097/JXX.0000000000000234.
[100] M.A. Orellana-Barrios, D. Payne, Z. Mulkey, K. Nugent, Electronic cigarettes - a narrative review for clinicians, Am. J. Med. 128 (2015) 674–681, https://doi.org/10.1016/j.amjmed.2015.01.033.
[101] Results from the Annual National Youth Tobacco Survey | FDA, (n.d.). ⟨https://www.fda.gov/tobacco-products/youth-and-tobacco/results-annual-national-youth-tobacco-survey⟩ (accessed July 16, 2024).
[102] J.Y. Chien, Y.C. Gu, C.H. Liu, H.M. Tsai, C.N. Lee, A.C. Yang, J. Huang, Y.L. Wang, J.K. Wang, C.H. Lin, Rapid detection of nicotine and benzoic acid in e-liquids with surface-enhanced Raman scattering and artificial intelligence-assisted spectrum interpretation, J. Pharm. Biomed. Anal. 233 (2023), https://doi.org/10.1016/J.JPBA.2023.115456.
[103] A. Alavalapadu, R. Mattamal, Vaping associated pulmonary injury, Int J. Integr. Pedia Environ. Med. 7 (2022) 8–12, https://doi.org/10.36013/ijipem.v7i.75.
[104] A. Kishimoto, D. Wu, D.F. O'Shea, Forecasting vaping health risks through neural network model prediction of flavour pyrolysis reactions, 2024 14:1, Sci. Rep. 14 (2024) 1–14, https://doi.org/10.1038/s41598-024-59619-x.
[105] N.C. Atuegwu, E.M. Mortensen, S. Krishnan-Sarin, R.C. Laubenbacher, M.D. Litt, Prospective predictors of electronic nicotine delivery system initiation in tobacco naive young adults: a machine learning approach, Prev. Med. Rep. 32 (2023), https://doi.org/10.1016/J.PMEDR.2023.102148.
[106] J. Shi, R. Fu, H. Hamilton, M. Chaiton, A machine learning approach to predict e-cigarette use and dependence among Ontario youth, Health Promot Chronic Dis. Prev. Can. 42 (2022) 21, https://doi.org/10.24095/HPCDP.42.1.04.
[107] A.L. Vázquez, C.M. Navarro Flores, B.H. Garcia, T.S. Barrett, M.M. Domenech Rodríguez, An ecological examination of early adolescent e-cigarette use: a machine learning approach to understanding a health epidemic, PLoS One 19 (2024), https://doi.org/10.1371/JOURNAL.PONE.0287878.
[108] I. Singh, V. Valavil Punnapuzha, N. Mitsakakis, R. Fu, M. Chaiton, A machine learning approach reveals distinct predictors of vaping dependence for adolescent daily and non-daily vapers in the COVID-19 era, 2023, Vol. 11, Page 1465, Healthcare 11 (2023) 1465, https://doi.org/10.3390/HEALTHCARE11101465.
[109] R. Fu, J. Shi, M. Chaiton, A.M. Leventhal, J.B. Unger, J.L. Barrington-Trimis, A machine learning approach to identify predictors of frequent vaping and vulnerable californian youth subgroups, Nicotine Tob. Res. 24 (2022) 1028–1036, https://doi.org/10.1093/NTR/NTAB257.
[110] T.T.T. Le, Key risk factors associated with electronic nicotine delivery systems use among adolescents, JAMA Netw. Open 6 (2023) E2337101, https://doi.org/10.1001/JAMANETWORKOPEN.2023.37101.
[111] L. Hassan, M. Elkaref, G. de Mel, I. Bogdanovica, G. Nenadic, Text mining tweets on e-cigarette risks and benefits using machine learning following a vaping related lung injury outbreak in the USA, Healthc. Anal. 2 (2022) 100066, https://doi.org/10.1016/J.HEALTH.2022.100066.
[112] Y. Ren, D. Wu, A. Singh, E. Kasson, M. Huang, P. Cavazos-Rehg, Automated detection of vaping-related tweets on twitter during the 2019 EVALI outbreak using machine learning classification, Front Big Data 5 (2022) 770585, https://doi.org/10.3389/FDATA.2022.770585/BIBTEX.
[113] Z. Xie, S. Deng, P. Liu, X. Lou, C. Xu, D. Li, Characterizing anti-vaping posts for effective communication on instagram using multimodal deep learning, Nicotine Tob. Res. 26 (2024) S43–S48, https://doi.org/10.1093/NTR/NTAD189.
[114] N. Shah, M. Nali, C. Bardier, J. Li, J. Maroulis, R. Cuomo, T.K. Mackey, Applying topic modelling and qualitative content analysis to identify and characterise ENDS product promotion and sales on Instagram, Tob. Control 32 (2023) E153–E159, https://doi.org/10.1136/TOBACCOCONTROL-2021-056937.
[115] S. Visweswaran, J.B. Colditz, P. O'Halloran, N.R. Han, S.B. Taneja, J. Welling, K. H. Chu, J.E. Sidani, B.A. Primack, Machine learning classifiers for twitter surveillance of vaping: comparative machine learning study, J. Med. Internet Res. 22 (2020) e17478, https://doi.org/10.2196/17478.
[116] G. Kostygina, H. Tran, L. Czaplicki, S.N. Perks, D. Vallone, S.L. Emery, E.C. Hair, Developing a theoretical marketing framework to analyse JUUL and compatible e-cigarette product promotion on Instagram, Tob. Control 32 (2023) E192–E197, https://doi.org/10.1136/TOBACCOCONTROL-2021-057120.
[117] G. Kong, A.S. Schott, J. Lee, H. Dashtian, D. Murthy, Understanding e-cigarette content and promotion on YouTube through machine learning, Tob. Control 32 (2023) 739–746, https://doi.org/10.1136/TOBACCOCONTROL-2021-057243.
[118] D. Murthy, R.R. Ouellette, T. Anand, S. Radhakrishnan, N.C. Mohan, J. Lee, G. Kong, Using computer vision to detect E-cigarette content in TikTok videos, Nicotine Tob. Res. 26 (2024) S36–S42, https://doi.org/10.1093/NTR/NTAD184.
[119] J. Vassey, C.J. Kennedy, H.C.H. Chang, A.S. Smith, J.B. Unger, Scalable surveillance of E-cigarette products on instagram and TikTok using computer vision, Nicotine Tob. Res. 26 (2024) 552–560, https://doi.org/10.1093/NTR/NTAD224.
[120] R. Lakatos, P. Pollner, A. Hajdu, T. Joó, A multimodal deep learning architecture for smoking detection with a small data approach, Front Artif. Intell. 7 (2024) 1326050, https://doi.org/10.3389/FRAI.2024.1326050/BIBTEX.
[121] N. Nargis, A.K.M.G. Hussain, S. Asare, Z. Xue, A. Majmundar, P. Bandi, F. Islami, K.R. Yabroff, A. Jemal, Economic loss attributable to cigarette smoking in the USA: an economic modelling study, Lancet Public Health 7 (2022) e834–e843, https://doi.org/10.1016/S2468-2667(22)00202-X.
[122] Z. Yuan, R. Zhou, H. Wang, L. He, Y. Ye, L. Sun, ViT-1.58b: Mobile Vision Transformers in the 1-bit Era, (2024). ⟨https://arxiv.org/abs/2406.18051v1⟩ (accessed December 3, 2024).
[123] B. Lim, S. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, Int. J. Forecast 37 (2019) 1748–1764, https://doi.org/10.1016/j.ijforecast.2021.03.012.
[124] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms, (2017). ⟨https://arxiv.org/abs/1707.06347v2⟩ (accessed December 3, 2024).

[125] M. Sewak, Deep Q Network (DQN), Double DQN, and Dueling DQN, Deep Reinf. Learn. (2019) 95–108, https://doi.org/10.1007/978-981-13-8285-7_8.

[126] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H.W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S.P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S.S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N.S. Keskar, T. Khan, L. Kilpatrick, J.W. Kim, C. Kim, Y. Kim, J.H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C.M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S.M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de A.B. Peres, M. Petrov, H.P. de O. Pinto,Michael, Pokorny, M. Pokrass, V.H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F.P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M.B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J.F.C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J.J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, G.P.T.-4 Technical Report, (2023). ⟨https://arxiv.org/abs/2303.08774v6⟩ (accessed December 3, 2024).

[127] J. Li, T. Sawaragi, Y. Horiguchi, Introduce structural equation modelling to machine learning problems for building an explainable and persuasive model, SICE J. Control Meas. Syst. Integr. 14 (2021) 67–79, https://doi.org/10.1080/18824889.2021.1894040.