Master thesis

Master of Science (M.Sc.)
Department of
Major : Data Science

Topic:

Unveiling Multifaceted Nature of Smoking: A Data Science Approach Utilising Public Health Datasets

Designing a Stacking Ensemble Framework for Smoking Behavior Prediction with Focus on Model Interpretability

Author: Joshil Fernandes

Matriculation Number: 86334288
First supervisor: Prof. Dr. Eman Abhukhousa
Second supervisor: Prof. Dr. Iftikhar Ahmed
Submitted on: 28.08.2024

Statutory Declaration:

I hereby declare that I have developed and written the enclosed Master Thesis completely by myself and have not used sources or means without declaration in the text. I clearly marked and separately listed all the literature and all the other sources which I employed when producing this academic work, either literally or in content. I am aware that the violation of this regulation will lead to failure of the thesis.

Berlin, 28.08.2024

. . . .

(Joshil Fernandes)

University of Applied
Sciences Europe
Iserlohn · Berlin · Hamburg

EIGENSTÄNDIGKEITSERKLÄRUNG / STATEMENT OF AUTHORSHIP

Joshil
Name | Family Name

Fernandes
Vorname | First Name

86334288
Matrikelnummer | Student ID Number

Unveiling Multifaceted Nature of Smoking
Titel der Examsarbeit | Title of Thesis

Ich versichere durch meine Unterschrift, dass ich die hier vorgelegte Arbeit selbstständig verfasst habe. Ich habe mich dazu keiner anderen als der im Anhang verzeichneten Quellen und Hilfsmittel, insbesondere keiner nicht genannten Onlinequellen, bedient. Alles aus den benutzten Quellen wörtlich oder sinngemäß übernommen Teile (gleich ob Textstellen, bildliche Darstellungen usw.) sind als solche einzeln kenntlich gemacht.

Die vorliegende Arbeit ist bislang keiner anderen Prüfungsbeh.rde vorgelegt worden. Sie war weder in gleicher noch in ähnlicher Weise Bestandteil einer Prüfungsleistung im bisherigen Studienverlauf und ist auch noch nicht publiziert. Die als Druckschrift eingereichte Fassung der Arbeit ist in allen Teilen identisch mit der zeitgleich auf einem elektronischen Speichermedium eingereichten Fassung.

With my signature, I confirm to be the sole author of the thesis presented. Where the work of others has been consulted, this is duly acknowledged in the thesis' bibliography. All verbatim or referential use of the sources named in the bibliography has been specifically indicated in the text.

The thesis at hand has not been presented to another examination board. It has not been part of an assignment over my course of studies and has not been published. The paper version of this thesis is identical to the digital version handed in.

27-08-2024 Berlin
Datum, Ort | Date, Place

Joshil
Unterschrift | Signature

Abstract

The global explosion in smoking-related disorders emphasizes how urgently creative detection and monitoring methods are needed. Conventional approaches, which mostly rely on self-reported data prone with errors and prejudices, typically miss the complexity of smoking behavior. Recent developments in wearable technology and machine learning offer fresh chances for real-time, objective data analysis-based solution addressing of these issues. Still, current studies have not completely utilized the possibilities for smoking detection by combining several bio-signals with sophisticated machine learning methods. This discrepancy in studies drives the present work to create a comprehensive predictive model analyzing bio-signals for smoking behavior by use of ensemble machine learning techniques. This work trains several machine learning models including logistic regression, random forest, and advanced ensemble approaches such gradient boosting and stacking classifiers using a mix of heart rate, blood pressure, and electrodermal activity data. The approach stresses thorough data preparation, feature engineering, and meta-learning framework implementation to improve prediction accuracy. With better accuracy and resilience in forecasting smoking status, the ensemble models—especially the stacking classifier—show clearly how better than conventional single-model techniques. Enhancing the prediction power of the model has proved mostly dependent on the integration of several bio-signals, therefore offering a richer, more accurate representation of smoking habit. The implications of this research are vast, suggesting a shift towards more reliable and real-time smoking detection methods that could be integrated into public health strategies and personal health monitoring systems. This study not only advances the field of health informatics but also sets a precedent for future research in behavioral prediction and monitoring using bio-signals. Finally, the creation of this predictive model represents a major advance in the approaches applied for smoking detection, so improving the outcomes in public health projects and individualized healthcare management. To completely realize its advantages, more investigation should look at the integration of other bio-signals and the implementation of this model in practical situations.

Contents

List of Figures

List of Tables

1 Introduction

The worldwide load of smoking-related diseases is among the most important public health concerns of the modern era. With around 8 million fatalities annually connected to smoking-related causes, tobacco use is a main cause of unnecessary death worldwide [Wor23]. Notwithstanding major public health campaigns and smoking cessation programs, the addictive nature of nicotine and the general availability of tobacco products have contributed to extend this epidemic. Mostly depending on self-reported data and clinical assessments, conventional methods of smoking detection and monitoring have not been able to fully handle the complexity of smoking behavior. Many times, these strategies are prone to underreporting, mistakes, and biases that hinder good intervention and policy-making.

Advances in wearable technology and machine learning have lately opened new directions for health monitoring and behavior detection[BLY^{+}18]. These technologies could revolutionize the identification and control of smoking by means of more exact, objective, real-time data. This work aims to leverage these technical advancements to generate a robust machine learning model capable of predicting smoking status based on bio-signal data. By incorporating multiple physiological markers—heart rate, blood pressure, and electrodermal activity—into a predictive model, this work seeks to give a more consistent substitute for standard smoking detection methods. Among other fields, this work has broad implications for public health surveillance, tailored medication, and smoking cessation campaigns.

1.1 Motivation

The urgent need for more effective responses and the increasing knowledge of the limits of conventional smoking detection methods inspire this work. Apart from being a significant cause to numerous chronic diseases including lung cancer, cardiovascular disease, and chronic obstructive pulmonary disease (COPD), secondhand smoke exposure seriously undermines public health[Cen23]. Although everyone is aware of these risks, smoking is still somewhat common in part that it is difficult to exactly monitor and identify smoking behavior.

Conventional methods of smoking detection include self-reported questionnaires and clinician interviews have challenges. These methods rely on people's honesty and recall, which could lead to considerable underreporting of smoking behavior[VMH^{+}20]. Moreover, these methods are often retrospective, which means they rely on people remembering their behavior over a certain period and so produce recall bias. The subjective nature of these exams can result in erroneous data tough to utilize for reasonable intervention strategies.

An interesting alternative for more traditional methods is wearable technology. Wearable devices can continuously monitor many physiological indicators, therefore offering real-time data that can be used to detect changes related with smoking behavior. For instance, smoking has been shown to cause changes in breathing patterns and skin conductance as well as sharp increases in blood pressure and heart rate. Analyzing these bio-signals using machine learning methods enables one to develop a predictive model quite accurate in spotting smoking events. This approach not only eliminates the restrictions of self-reported data but also provides a non-invasive, real-time monitoring tool accessible in both clinical and daily surroundings.

Furthermore, a big advancement in personalized medicine is the integration of machine learning into systems of health monitoring. Capabilities of machine learning techniques are large datasets and pattern recognition that might not be immediately clear to human observers. In the context of smoking detection, this suggests that machine learning models can maybe identify minute physiological changes that precede smoking events, therefore facilitating early intervention and support for people seeking to quit smoking. This research is motivated by the possibility to exploit these technical improvements to build a more effective instrument for smoking detection; the ultimate goal is to reduce the global burden of tobacco-related diseases by means of this technology.

Moreover, several research proving the effectiveness of bio-signals in several applications connected to health support the increasing interest in using these approaches for behavioral detection[SVA+22]. From stress and anxiety to sleep disorders and metabolic diseases, research have shown that bio-signals can be used to identify a wide spectrum of behaviors and situations[JPK+14] [CFN+21]. This paper investigates the specific application of bio-signal analysis in the framework of smoking detection, therefore extending on these findings. The motivation to look at this specific use comes from the ability to assist a critical field of public health and encourage individuals in their endeavors to stop smoking.

1.2 Objective

The primary objective of this work is to develop a machine learning model based on bio-signal data adept of exactly recognizing smoking behavior. This aim is driven by the need to surpass the constraints of present smoking detection methods and provide a more objective and trustworthy instrument for identifying smoking events. Combining many bio-signals—heart rate, blood pressure, and electrodermal activity—into a predictive model suited for both personal health monitoring and mass public health campaigns helps the study to attain this goal.

Reaching this target will cause the study to focus on several crucial tasks. First it will include compiling and preprocessing bio-signal data from a diverse population source. This will ensure that the model develops on a wide range of physiological data, hence enhancing its generalizability and accuracy. Second, several machine learning approaches will be created and evaluated to find the most effective way for prediction of smoking habit. This will constitute the evaluation of several model topologies and training strategies in order to maximize the performance of the model. At final, the research will be based on exhaustive testing on both training and validation sets, so confirming the dependability and correctness of the model.

A secondary aim of this effort is to identify the most significant bio-signals associated with smoking behavior. By means of the investigation of the contributions of numerous physiological markers to the predictions of the model, the work aims to clarify the basic mechanisms of smoking behavior and identify potential biomarkers relevant in next research and clinical practice. Apart from improving the accuracy of the forecast model, this objective allows us to progress our knowledge of the physiological effects of smoking.

Ultimately, the project aims to provide a tool to support both public health campaigns and individual efforts at smoking cessation. By providing a consistent and objective method for identifying smoking habit, the study has the chance to reduce the incidence of smoking-related diseases and boost the efficacy of smoking cessation programs. In the field of health informatics, the efficient achievement of these objectives will be significantly valued and provide a foundation for next research in this field.

1.3 Problem Statement

Effective detection of smoking behaviors remains a challenging and unresolved issue even with considerable advancements in healthcare technology. Mostly depending on self-reported data, present methods are prone to mistakes and prejudices that lead to a considerable underestimation of smoking prevalence and therefore a failure to effectively target and help those most at risk. Big public health efforts where measuring the performance of policies aimed to reduce smoking rates depends on specific and reliable statistics intensify this issue.

Even more aggravating the problem is the lack of integration among numerous bio-signals that would provide a more whole picture of a person's smoking behavior. Most existing models limit their application over different populations and accuracy by concentrating on specific physiological markers. This work solves these constraints by means of a machine learning model that not only incorporates numerous bio-signals but also employs ensemble learning methodologies to boost the prediction capability of the model.

1.4 Thesis Scope

With a deliberately broad scope, this thesis addresses the evolution, validation, and probable application of a machine learning model for smoking behavior detection using bio-signal data. This study covers many significant fields: the collecting and preprocessing of bio-signal data, the construction of a prediction model using multiple machine learning approaches, and the performance evaluation of this model in many settings. The study is supposed to be exhaustive such that the produced model is not only theoretically solid but also practicable and scalable for real usage.

Still, the research also focus in some rather particular areas. It limits its research especially to bio-signals—that is, skin conductance, blood pressure, and heart rate—that wearable sensors can regularly collect. This focus ensures that the findings of the study directly relate to the fast growing field of wearable health technologies. The thesis provides a strong platform for next research aiming at more comprehensive exploration of this element even though it does not contain the clinical application of the model. Through an emphasis on the technical and methodological challenges of model construction and validation, this thesis aims to significantly progress the fields of health informatics and public health.

1.5 Purpose and goal

This paper aims to examine smoking activity detection using bio-signal analysis-based machine learning methods. Being a main cause of avoidable diseases, smoking requires innovative approaches of monitoring and intervention. By building a machine learning model that can properly identify smoking behaviors based on bio-signals, our effort seeks to provide a tool that can be used in both public and individual health environments, thereby boosting the efficacy of smoking cessation programs and public health campaigns.

This effort attempts to generate a scalable, non-invasive, reliable instrument appropriate for inclusion into present health monitoring systems. Apart from improving the detection of smoking patterns, this tool would enable the development of more targeted and successful treatments. By offering a novel and potent approach of health monitoring, the study intends to greatly help in the fight against smoking-related diseases by reaching this target.

1.6 Outline

The carefully constructed structure of this thesis is intended to enable the reader to negotiate the difficult process of building a machine learning model for smoking behavior detection with bio-signal data.
There are seven chapters to the thesis, each covering a vital facet of the investigation.

Chapter 2 offers a thorough review of the theoretical underpinnings required to grasp the complexity of behavioral pattern prediction, especially in the field of smoking behavior employing bio-signals. This chapter addresses the application of several machine learning models including logistic regression, random forests, and ensemble approaches in the study of bio-signals to improve the prediction accuracy and resilience of smoking behavior detection. Important ideas and techniques such feature engineering and model interpretability—qualities necessary for efficiently adding bio-signals into predictive models—also are discussed. This chapter builds a critical theoretical framework supporting the later empirical study by means of thorough definitions and descriptions of important words and methodologies.

Chapter 3 examines earlier research on behavioral prediction using machine learning models, with an eye toward smoking detection especially. Deep learning and ensemble models as well as other machine learning techniques are covered in this chapter together with their success in predicting smoking habit and associated health effects. It explores the intricacies of how several models—including SVM, Random Forest, and deep learning—have been used to examine bio-signals associated with smoking in addition to forecast behavior. The chapter also looks at methods and difficulties in feature engineering and data preparation meant to improve model performance. Furthermore discussed are techniques applied to manage data set imbalances and enhance forecast accuracy by means of advanced feature selection approaches.

Chapter 4 outlines the fundamental research issues and problem statement guiding the study of leveraging bio-signals for machine learning-based smoking behavior prediction. The chapter covers the shortcomings of conventional detection techniques and introduces the use of bio-signal data to improve forecast dependability and accuracy. Two key research questions are raised by it: the first looks at bio-signal analysis of several different machine learning models, including XGBoost and logistic regression. The second issue evaluates if accuracy of a stacking classifier incorporating several models beats that of single-model techniques. The implementation of advanced machine learning methods to enhance public health interventions against smoking depends critically on this chapter. .

Chapter 5 describes the model's implementation approach and emphasizes the significant features of the coding and algorithmic choices made during development. The difficulties faced during implementation—such as the management of missing data and the optimization of model parameters—as well as the techniques used to solve them—this chapter also addresses The chapter offers a detailed manual for the implementation procedure, therefore guaranteeing that the research can be repeated and expanded by next generations of researchers.

Chapter 6 concentrates on the validation and assessment of the produced model. It details the testing surroundings, the evaluation criteria applied, and the findings of many tests. The chapter also covers the scalability of the model, its performance over several demographic groups, and the possibility to include the model into already in use health monitoring systems. The thorough assessment guarantees that the model is dependable, strong, and relevant for practical problems.

Chapter 7 summarizes the main conclusions, addresses the limits of the research, and offers suggestions for next projects, therefore closing the thesis. This chapter also considers the possible influence of the studies on public health campaigns and smoking cessation programs, therefore stressing the contributions made by the research to the domains of public health informatics and public health itself. The chapter offers a road map for next studies, pointing up topics requiring more inquiry and suggesting fresh paths of application of machine learning in health monitoring.

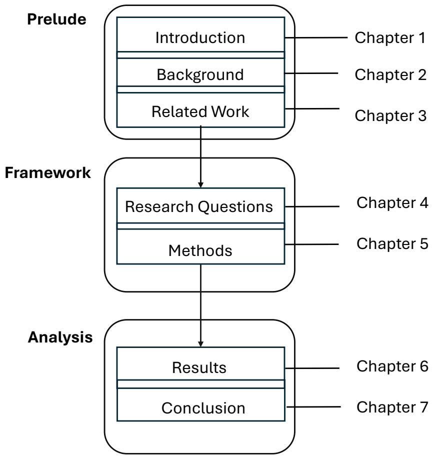

Figure 1.1: Thesis Structure and Flow of reading

2 Theoretical Background

2.1 Background

Accurate behavioral pattern prediction—such as smoking habits—is absolutely essential for preventative care and tailored treatment plans in the new terrain of healthcare analytics. In this field, bio-signal analysis is very important since it provides thorough understanding of physiological states by means of data acquired from sources. Logistic regression and random forests among other classic machine learning (ML) models have been extensively used to examine these bio-signals for smoking behavior prediction. These algorithms have shown great accuracy in identifying bio-signal data depending on smoking-related physiological changes and shine in pattern identification. Under such circumstances, Caiafa et al. (2020) address sophisticated decomposition techniques that efficiently handle the problems presented by small, partial, or noisy datasets, hence improving the prediction capacities of machine learning models[CSCMP$^+$20]. These traditional models can be limited in real-world applications where data may be sparse or noisy, though, since they sometimes depend on large datasets to run well. By aggregating several algorithms to increase prediction accuracy and model robustness, ensemble learning methods such stacking, boosting, and bagging have become increasingly effective substitutes for single models, therefore resolving their constraints [LLQ$^+$20] [HELMGA20]. These methods not only improve the general performance but also give a more complete knowledge of the intricate interactions in bio-signal data by using the special strengths of every model in an ensemble. This work is to investigate the efficiency of these advanced machine learning approaches in enhancing the prediction of smoking behaviors, thereby supporting more efficient public health campaigns and customized treatment strategies.

2.1.1 Definitions

Bio-signals: Bio signal can be any typical sampling of the human body that is useful to interpret the physiological state it is in. There are now a countless number of bio signals you could work on from an Electromyograms (EMG), Electrocardiograms(ECG), and to have it even worse the Electroencephalogram(EEG). This all information when studies correctly can tell this much of things about the human health which is gonna help a lot in research. This thesis uses Biosignals as a primary data source for prediction.

Exploratory Data Analysis (EDA): A vital first stage in data preparation, exploratory data analysis (EDA) summarizes the essential features of a dataset usually using visual approaches. EDA is meant to help one test theories, recognize trends, comprehend the structure of the data, and spot abnormalities. Often employed during EDA to find underlying patterns and relationships in the data, techniques such histograms, scatter plots, box plots, and correlation matrices can direct next modeling decisions.

Machine learning (ML): Within artificial intelligence (AI), machine learning is a subset whereby algorithms enable computers to learn from data and generate predictions or judgments. When examining complicated bio-signals where conventional statistical methods might not be sufficient, machine learning approaches are very successful in identifying patterns inside big datasets. In the framework of this study, machine learning models are trained to recognize patterns in bio-signal data corresponding with smoking occurrences, hence detecting smoking habit.

Ensemble Models: Ensemble models are a class of machine learning model whereby predictions from several base models are aggregated to raise general performance. Under ensemble learning, the theory is that the final prediction is more accurate and robust than that of any one model by aggregating the strengths of several models. Common ensemble methods consist on bagging, boosting, and stacking. In this thesis, smoking habit using bio-signal data is improved in prediction accuracy using ensemble models [DYC^{+}20].

Boosting Models: Boosting is a sequence of models whereby each new model fixes mistakes caused by the last one. Combining the models generates a strong learner at last. By concentrating on challenging scenarios that past models failed with, boosting algorithms as XGBoost, LightGBM, and CatBoost are especially successful in increasing model accuracy. This thesis improves the prediction capacity of the machine learning models for smoking behavior detection by use of boosting models.

One-Hot Encoding: One-hot encoding is a technique for turning categorical variables into a format fit for machine learning systems. Every category in a categorical feature is converted into a new binary feature whereby the lack of the category is denoted by 0 and its presence by 1. One-Hot Encoding would translate, for instance, a categorical variable "Color" with three categories—"Red," "Blue," and "Green"—into three distinct binary features, one for each color. In machine learning models, managing categorical data calls for this approach. Hancock and Khoshgoftaar underline in their survey the crucial part of methods like one-hot encoding in properly preparing categorical data for neural networks, therefore stressing their relevance over many data-driven applications [HK20].

Feature Engineering: In feature engineering—the process of choosing, altering, or generating new features (variables) from raw data to enhance machine learning model performance—variables Improving the prediction ability of models depends on good feature engineering, particularly in cases involving complicated data such as bio-signals. This procedure ensures that the most useful features of the raw bio-signal data are caught to enhance model accuracy by converting them into a more fit format for model training.

Heart Rate Variability (HRV): HRV is a basic physiological indicator applied in this work as a characteristic in smoking detection. Under control of the autonomic nervous system, HRV is the variation in the time interval separating consecutive heartbeats. Among the various physiological and psychological diseases connected to changes in HRV are stress, anxiety, and drug use. Within the framework of smoking detection, HRV provides interesting analysis of the body's reaction to nicotine ingestion, so acting as a relevant bio-signal for the built machine learning models in this thesis.

Interpretability Models: Tools and approaches used to explain and grasp the predictions generated by sophisticated machine learning models are interpretability models. Local Interpretable Model-agnostic Explanations (LIME) is applied in this thesis as an interpretability model to clarify the inner operations of ensemble models. LIME helps to discover which features are most important in the decision-making process by approximating the complicated model with a simpler, interpretable model (such a linear model) around each prediction, hence enabling explanations. In healthcare, where knowledge of the justification behind model predictions can affect clinical decisions, this openness is absolutely vital [VSM24].

Stacking: In ensemble learning, stacking is a method whereby several models—usually of different kinds—are trained and integrated by another model, also known as a meta-learner to get a final prediction. The base models learn from the dataset; the meta-learner is trained on the predictions of the base models, therefore hopefully enhancing the predictive performance by lowering overfitting and capturing many facets of the data.

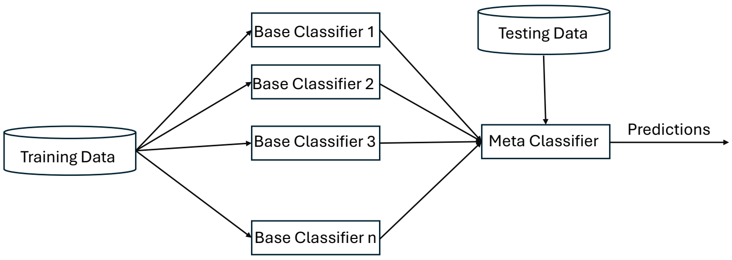

Figure 2.1: Stacking Ensemble Overview

Standardization: A method of data preparation known as standardization rescales features such that their mean is zero and their standard deviation is one. In machine learning especially when features have variable units or scales, this is especially crucial since it guarantees that the model does not give any one feature undue weight depending just on their scale. In geological data interpretation, Corbett et al. also address the crucial relevance of up-scaling and cross-scaling, which modify measurements to meaningful scales and compare several kinds of data, so stressing the universal relevance of suitable scaling in data analysis [CJS98].

2.2 Existing Methods

Using the strengths of several machine learning models, ensemble stacking has become a potent method in the field of medical diagnostics in recent years that greatly increases predictive accuracy. Many research studies have effectively used ensemble stacking techniques in medical environments, stressing their applicability to biosignal-based smoking behavior prediction.

2.2.1 Ensemble Stacking with Multi-Objective Optimization

Using an ensemble stacking technique combined with multi-objective optimization, the authors of "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," created a novel machine learning framework. Their approach, NSGA-II-Stacking, chose an ideal ensemble of base classifiers using a non-dominated sorting genetic algorithm (NSGA-II), then integrated via a meta-classifier to predict Type-2 diabetes mellitus. Against conventional machine learning models, this method showed notable gains in classification accuracy, sensitivity, specificity, and other performance criteria. Using biosignals, the NSGA-II-Stacking method attained an accuracy of 83.8%, sensitivity of 96.1%, and an AUC of 85.9%, so highlighting the potential of ensemble stacking in enhancing the robustness and accuracy of predictive models, which could be effectively applied to smoking behavior prediction [SS20].

2.2.2 Hybrid Stacked Ensemble with Genetic Algorithms

In their paper "Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction," Jafar Abdollahi and Babak Nouri-Moghaddam underlined in the context of diabetes prediction the efficiency of a hybrid stacked ensemble model. This work highly accurately diagnosed and predicted the start of diabetes by combining genetic algorithms with ensemble learning methods. Reaching an accuracy of 99%, the "Stacked Generalization based Metaheuristics" system combined several machine learning models including artificial neural networks, SVMs, and decision trees into a strong prediction model. The genetic algorithm optimization applied in this work could be modified to enhance the feature selection process for biosignals, so improving the performance of the ensemble model for smoking behavior [ANM22].

2.2.3 Patient-Specific Stacking Models in Healthcare

El-Rashidy et al. (2020) addressed a better patient-specific stacking ensemble model for mortality prediction in intensive care units. Each tuned for different patient data modalities, this method combined logistic regression, k-nearest neighbors, decision trees, and multilayer perceptrons among other machine learning classifiers. With regard to accuracy, precision, recall, and AUC scores, our ensemble technique combined the decisions of several classifiers to generate a coherent forecast, much surpassing state-of-the-art methods. This method is especially pertinent for constructing models to use biosignals to forecast smoking behavior by means of a similar fusion of specialized classifiers, hence improving prediction accuracy [**?**].

2.2.4 Stacking Ensemble for Early Disease Detection

In his paper "Early Detection of Coronary Heart Disease Using Ensemble Techniques," Shorewala investigated the efficiency of bagging, boosting, and stacking ensemble techniques for early coronary heart disease (CHD). Combing K-Nearest Neighbors, Random Forest, SVM using logistic regression as the meta-classifier, the stacking ensemble showed the best accuracy at 75.1%. Highly relevant to predicting smoking behavior using biosignals, this study shows the power of stacking ensembles in synthesizing the strengths of many base classifiers to improve predictive performance in medical diagnostics [Sho21].

2.2.5 Stacking Models for Predicting Hospital Admissions

Using historical admissions data, air quality, and meteorological data, Hu et al. (2020) revealed an inventive stacking ensemble model used to predict daily hospital admissions for cardiovascular disorders. The model included XGBoost and random forest, among other machine learning techniques including linear regression, support vector regression, and tree-based models. Regarding accuracy and other criteria, the ensemble technique exceeded individual models by combining forecasts from several models. This approach fits very nicely with using ensemble stacking to combine several biosignal data for smoking behavior prediction, so enhancing possibly prediction accuracy and robustness [HQS$^+$20].

2.2.6 Stacking with LIME for Model Interpretability

Designed to forecast depression in Parkinson's disease sufferers, Nguyen and Byeon (2023) developed a LIME-based stacking ensemble model. Using Logistic Regression as the meta-model, the paper merged several machine learning techniques including LightGBM, K-Nearest Neighbors, Random Forest, Extra Trees, and AdaBoost. Especially, the model included LIME (Local Interpretable Model-Agnostic Explanations), so improving the transparency by clarifying forecasts produced by the otherwise "black-box" stacking ensemble model. LIME is a useful tool in clinical environments since its integration guarantees that the predictions of the model are interpretable and practical. Using biosignals, this framework might be modified to forecast smoking behavior, hence improving both accuracy and interpretability [NB23].

2.2.7 Advanced Stacking with SHAP for Medical Predictions

In their 2024 paper "A Stacking Ensemble Model for Predicting the Occurrence of Carotid Atherosclerosis," Zhang et al. showed how well a stacking ensemble model improved prediction accuracy for carotid atherosclerosis (CAS). Integrating several machine learning techniques—Logistic Regression, Random Forest, SVM, XGBoost, and Gradient Boosting Decision Tree—into a stacked ensemble framework, the study with an AUC of 0.893, this combined approach much exceeded individual models. With an eye toward the contribution of endocrine-related markers to CAS risk, the study also used the SHAP (SHapley Additive exPlanations) approach to interpret the model's predictions. This approach directly relates to estimating smoking behavior using biosignals by highlighting the potential of stacking ensemble models in challenging medical prediction problems [ZTW$^+$24].

All things considered, the body of current research on ensemble stacking in medical predictions emphasizes the great benefits of this method in managing challenging and varied datasets, raising predictive accuracy, and so strengthening model interpretability. Often combined with sophisticated methods like genetic algorithms or SHAP for model explanation, the integration of several machine learning algorithms through stacking highlights the resilience of these models in medical diagnosis. This piece of work offers a strong basis for applying comparable ensemble techniques employing biosignals to predict smoking habit, hence possibly producing more accurate and useful predictions in healthcare environments.

2.3 Evaluation metrics

The performance of the machine learning models in this work is evaluated under many criteria. These measures give a whole picture of the accuracy, precision, and general performance of the models in estimating smoking behaviors from bio-signals. Among the measures applied are accuracy, F1 score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC).

2.3.1 Accuracy

Among the overall number of cases investigated, accuracy is the percentage of accurate results—that is, both true positives and true negatives. It's a simple statistic showing how often the model accurately forecasts smoking status. But accuracy by itself might not be enough, particularly in situations when the class distribution is skewed since it does not differentiate between the several kinds of mistakes the model produces.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.1}$$

2.3.2 F1 Score

Particularly helpful in situations with imbalanced classes, the F1 Score is the harmonic mean of recall and accuracy. In these situations, it is a more useful statistic than accuracy since it strikes a mix between recall (the proportion of genuine positive results in all positive predictions) and precision—that is, the proportion of true positive results in all real positives. The F1 Score is calculated as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.2}$$

2.3.3 ROC AUC

The performance metric in classification problems for different threshold values Area Under the Receiver Operating Characteristic Curve(ROC AUC) A ROC curve (receiver operating characteristic) is a plot of the true positive rate against the false-positive rate, and AUC represents how well your model ranks between possible thresholds. An AUC value of 1 implies a perfect classification while greater values indicate better model performance.

$$\text{AUC} = \int_0^1 \text{TPR}(f)\, d(\text{FPR}(f)) \tag{2.3}$$

3 Related work

3.1 Machine Learning and Deep Learning Models for Behavioral Prediction

Recently, utilisation of machine learning algorithms to predict smoking behavior and the related health outcomes are getting huge attention. Lai et al.: developed various classification models, namely Artificial Neural Networks (ANN), random forests(RF) Support Vector Machines(SVM) and Logistic Regression(LoR). Methods: Two datasets are used in this study by Penzes et al.(2021) to predict smoking cessation outcomes, where ANN has the best performance (AUC = 0. Part Sixed): a smoking cessation program in Taiwan [LHCH21]. Comparably, Caccamisi et al. (2020) demonstrated that using a Support Vector Machine (SVM), combined with Natural Language Processing to classify patients' status smoking in their Electronic Medical Records, resulting in 98.14% accuracy and an F-score of 0.981[CJDR20]. Huang et al. These represented an even larger advancement of the field, especially because future work as hypothesized by Gold (2023) was used in these works and data detecting smoking related transcriptome abnormalities of blood samples [24]. Their study produced SVM models with high sensitivity to differentiate smoking status using feature selection methods such as Boruta and LASSO [HMR$^+$23]. Cho et al. (2020) proposed the use of Internet of Things (IoT)-based sensors and machine learning algorithms for indoor smoking detection, where a non-linear SVM model with 88% F1 score and an accuracy is revealed as a top-performing algorithm[Cho20].

Focusing on young, Choi et al. (2021) built nicotine addiction prediction models utilizing Random Forest and LASSO, revealing that Random Forest gave greater accuracy (73.42%) in predicting addiction among e-cigarettes and hookah users [CJF$^+$21]. In order to forecast smoking occurrences, Abo-Tabik et al. (2021) also integrated Control Theory with a Bagged Decision Tree, therefore attesting to the capacity of machine learning models to consider personal variations in smoking behavior with a high classification accuracy of 91.075% [AT21]. Ali et al. (2020) proposed a smart healthcare monitoring system based on ensemble deep learning and feature fusion, therefore extending the usage of machine learning in healthcare. Though their approach can be employed for smoking detection with the ensemble model attaining 98.5% accuracy, their concentration on heart disease limits this [AESI$^+$20]. With SVM obtaining an AUC of 98.7%, Thakur et al. (2022) developed a real-time smoking activity detection model using data from a wrist-worn Inertial Measurement Unit (IMU) sensor, so demonstrating the potential of integrating wearable technology with machine learning for just-in-time smoking cessation interventions[?].

3.2 Feature Engineering and Data Preprocessing for Behavioral Data

Improving the effectiveness of machine learning models for behavioral prediction—including smoking detection—depends critically on feature engineering and data preprocessing. Using a VGG-16 pretrained network, Nakayiza and Ggaliwango (2021) created an interpretable feature learning framework for smoking behaviour detection. The study concentrated on finding important characteristics, especially those pertaining to the mouth area and smoke, which were absolutely necessary in precisely spotting smoking behavior in pictures. Using Layer-wise Relevance Propagation (LRP), they demonstrated the relevance of particular characteristics in the decision-making process, therefore stressing the need of explainability in feature learning for behavioral data analysis [HM21]. Likewise, Davagdorj et al. (2019) investigated how class imbalance in smoking cessation intervention datasets might be addressed using the Synthetic Minority Over-sampling Technique (SMote). Emphasizing the need of resolving class imbalance in behavioral data to increase prediction accuracy, the researchers improved the precision and recall of many machine learning classifiers, including Naive Bayes and Random Forest, by increasing the representation of minority classes[DLPR20].

Yang et al. Mantellini and Soni (2020) reiterated this in the context of behavioral prediction challenges, reinforcing that a well-considered preprocessing with feature engineering is also a must-have. Their work addressed dataset imbalance in ovarian cancer detection using SMote in concert with Support Vector Machine SMOTE (SVMSMote). Although used to cancer diagnosis, these methods are equally relevant to behavioral prediction problems such as smoking detection, where class imbalance can dramatically affect model performance [YKS20]. To improve the prediction accuracy of thyroid disease, Chaganti et al. (2022) investigated numerous feature engineering techniques including forward feature selection, backward feature elimination, and machine learning-based feature selection utilizing an additional tree classifier). These methods enabled the most pertinent feature identification from big datasets, therefore enhancing the performance of models such as Random Forest with a 99% accuracy rate. This work emphasizes the need of choosing suitable features and the part improved feature selection techniques play in raising model accuracy in predicting activities connected to health [CRDLTD$^+$22].

To identify COVID-19 cases from chest X-ray pictures, Nasiri et al. (2022) suggested a new framework combining deep learning with the ANOVA feature selection approach. By essentially lowering the complexity of the feature space, this method lets the model concentrate on the most pertinent features. Highly relevant to behavioral prediction models like as those used for smoking detection, the integration of DenseNet169 for feature extraction and ANOVA for feature selection shown great classification accuracy and lowered computational cost[NA22]. Laatifi et al. (2021) underlined even more how well advanced feature engineering methods improve behavioral prediction. Their work on COVID-19 severity prediction used Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, ANOVA F-statistic ensemble (AFSE), and Mutual Information for feature selection.

These techniques substantially increased the predictive power of machine learning models, which implies their prospective applications in complex behavioral data sets for use like smoking detection[LDB$^+$22]. For Early Chronic Disease Prediction, Al-Jamimi (2024) Indicated a Mixed Synergistic Feature Engineering Technique with Ensemble Learning. The study achieved significant increases in prediction accuracy through Recursive Feature Elimination (RFE) using a Support Vector Machine (SVM), as well as refining characteristics on the dataset with an eXtreme Gradient Boosting(XGBoost)classifier. Finally, incorporating hyperparameter tuning using Bayesian optimization further boosted the performance of our models. This complex method also calls for complicated feature engineering and optimization strategy which is equally applicable to behavioral prediction models like smoking detection[AJ24].

3.3 Ensemble Models in Behavioral Prediction

Especially in difficult prediction tasks such as behavioral outcomes, ensemble learning has been proven to be very effective for increasing effectiveness and robustness of machine-learning models. Kalagotla et al., by employing many base learners (MLP, SVM and LR), (2021) proposed a new stacking model to predict diabetes. This is technique of stacking was designed to leverage the characteristics over multiple methods tog ether in order do well in prediction. The stacking model outperforms baseline ensemble approaches such as AdaBoost by achieving higher accuracy, precision and F1-score. For behavioral prediction applications, such smoking detection, where data complexity and variability are major obstacles, the model's capacity to manage the non-linearity of data and lower the impact of individual model shortcomings makes it a desirable method[KGG21].

In patients with acute coronary syndrome, Zheng et al. (2021) also created a stacking ensemble model to forecast significant negative cardiovascular events (MACE). Based on seven generally used machine learning algorithms—Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost, and AdaBoost—the model utilized as base learners combines seven With notable increases in precision, recall, F1-score, and AUC, the stacking ensemble model—especially when combined with the SMOTETomek hybrid sampling technique—perfomed better than individual models in handling imbalanced data and attained an amazing accuracy of 98.62%. This work emphasizes the ability of stacking ensemble techniques to raise predictive accuracy in fields connected to health-related issues, hence extending their use to behavioral prediction challenges including smoking detection [ZSL21].

Gollapalli et al. (2022), for instance, used a dataset from Saudi Arabia . A new stacking ensemble for the diagnosis of three diabetes mellitus conditions. The paper also incorporated Bagging K-NN, Bagging Decision Tree and a similar model by meta-classifier k-nn with 94.48% in accuracy & weighted recall as well demonstrating Cohen's kappa score is around 0.9172 The findings indicate stacking ensemble methods are able to deal with imbalanced data well and improve the prediction of complex health status, thus supporting their potential utility for behavioural predictions models including smoking detection[GAA^{+}22]. Chiu et al. (2022) conducted a research to predict the mortality of ICU patients with heart failure using various methods. An improved stacking ensemble model by Base classifiers were built using Random forest, Support vector classification and K-Nearest Neighbors along with Light Gradient Boosting Machine (LGBM), Bagging, AdaBoost as other machine learning techniques in an ensemble stacking model, which showed an accuracy of 95.25% and AUCROC performance at 82.55%. This work demonstrates how ensemble models could be designed to combine the strengths of multiple algorithms, and thus effectively improve predictive accuracy in key healthcare settings (and potentially extend these lessons towards behavioral prediction) for example smoking detection[CWC^{+}22].

Fatima et al.(2023) via a composite of base classifiers, such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA) and Nu-Support Vector Machine (Nsvc), Multilayer Perceptron(MLP) also suggested a Stacking Ensemble Machine Learning Algorithm (SEMLA), which dealt with heart disease prediction. The SEMLA model outperformed constituent models with 97.28% accuracy, 98.4% sensitivity and a specificity of 96.21%. The model also performed well in several metrics including F1-score, Matthews correlation coefficient (Mcc), and receiver operating characteristics-area under the curve (ROC-AUC). More broadly, in a similar vein with respect to challenging prediction tasks involving behavioral outcomes like smoking detection, this study underscores the benefits of stacking ensemble methods — allowing for different classifiers to be combined into one model whilst minimising overfitting and maintaining relative good performance[FKT^{+}23]. Ghasemieh et al.(2023) A new Stacking Ensemble Learner (SEL) model was introduced, With combination of multiple base classifiers including Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree(DT) etc, into ensemble learning in this SEL model using XGBoost as a meta learner. The goal of this approach is to retain high detection capability while reducing strong consistency, across multiple runs. The SEL model was trained with a comprehensive dataset containing extensive clinical and biomedical data—at the patient, population level—curated by more than 1 hundred Medical Information Technology Inc., (MIT) Laboratory for Computational Physiology. It also outperformed conventional single-based classifiers in tasks that are crucial to medical predictive activities with high accuracy(88%) and strong recall, F1 scores — demonstrating the effectiveness of SEL model. The study suggests the importance of this model in cases where high reliability is needed, such as predicting readmission to follow-up care (or behavioral prediction tasks like smoking detection)[GLB+23].

The work of Mohapatra et al.(2023), that follows a two level ensemble using stacking classifiers with many machine learning techniques, eveloped a Heart Disease Prediction Model. In the meta-level they trained a Multilayer Perceptron to aggregate the base models outputs, and at the base level where different classifiers were used as Random Forest (RF), Multilayer Perceptron (MLP), K-Nearest Neighbors(KNN), Extra Trees(ET), Extreme Gradient Boosting(XGI). Stacked model scored 91.8% and had a sensitivity of 92.6%. The integrated classifier demonstrated substantially enhanced predictive performance over individual classifiers, especially in working with difficult medical datasets, indicating that this combined technique has the potential to be utilized for behavioral prediction applications such as smoking detection[MMPM23]. The study of Davagdorj et al.(2022) was carried out in South Korea and United States using the data from National Health and Nutrition Examination Survey (NHANES) datasets. In the present version, they proposed a XGBoost-dependent strategy for forecasting SiNCDs associated with cigarette smoking. This hybrid feature selection approach (HFS) used in the study included three stages: t-test and chi-square tests for initial features extraction, multicollinearity analysis to remove correlated features and finalities showcase from Least Absolute Shrinkage Selection Operator(LASSO). The chosen features were then fed to the XGBoost model which outperformed many of our baseline models Logistic regression, Random forest and KNN The accuracy, sensitivity and specificity demonstrated by the framework could be important for an early diagnosis of SiNCDs necessary to set aggressive preventive strategies. This approach demonstrates a promising result of ensemble learning with advanced feature selection methods in challenging health prediction applications and the behavior prediction such as smoke detection may be inferred from this method[DPTUR20].

Huang et al (2022) predicted the cardiovascular risk based on wearable device data and lifestyle characteristics with ensemble machine learning methods. Ensemble with Naive Bayes, Random Forests, and Support Vector Classifier for low risk predictions, Ensemble with generalized linear regression model, SVM,Stochastic Gradient Descent Regression-High Risk categories were used to build them, the researchers trained these models on a Southeast Asian cohort across variables that featured detailed lifestyle questionnaires as well as continuous blood pressure monitoring. The ensemble models surpassed the conventional Framingham Risk Score (FRS) with an AUC of 0.791 for low-risk groups and 0.790 for high-risk group This study demonstrates how an ensemble learning approach combines multiple data feeds to improve model accuracy, a method that could be useful for smoking identification prediction tools and other behavioral detection applications[HYC+22]. By leveraging diverse algorithmic capabilities and effectively managing data heterogeneity, this combined study highlights the seminal roles of ensemble learning models in improving the prediction for complex traits such as smoking.

4 Research Questions

Detecting and tracking smoking efforts using bio-signals are a big step in the field of public health technology toward handling an obvious encumbrance from numerous diseases driven by cigarette use. Several prior research results reported that machine learning especially ensemble models could improve the quality of prediction behavior with a good performance to overcome changes in environment. The inclusion of bio-signals (e.g., heart rate, blood pressure and other physiological markers) into predictive models allows to predict smoking in a non-invasive way; especially since self-reported data are liable more biased rates. This study differs from prior work in that we focused on using ensemble learning models to increase the accuracy of bio-signal measurement for predicting smoking behavior. We aim to accomplish this by conducting a comprehensive investigation of numerous machine learning techniques and developing an accurate predictive tool that not only improves detection accuracy, but also provides interpretable information on the influential factors associated with smoking behavior for more effective public health interventions.

4.1 Problem Statement

As a recognition of the worldwide consequences related to smoking disorders, an interest has been increased on developing more sophisticate schemes for monitoring and differentiate patterns in cigarette use. However, using traditional methods (e.g., self-reported data), a more connected or real-time monitoring is still not possible and methodological bias caused by the reporting of inaccurate detail have been repeatedly raised[OCP+20] [FKM+23]. These problems demonstrate the need for better, unbiased and real-time ways to detect smoking behaviors that can support public health efforts as well as strategies in more personalized healthcare. One potential alternative to self-reports is the incorporation of bio-signal data (e.g., heart rate, blood pressure and respiratory patterns), which are continuous, objective measures that can quantify smoking behaviour more reliably. However, this type of data is quite challenging to use in predictive models. The complexity and highly dynamic nature of bio-signals, as well the necessity for real-time processing require advanced analytical tools which can effectively represent these physiological markers.

Although machine learning models for predictive analytics have been promising, there is minimal literature on the optimization and synthesis of these models with bio-signal data pertaining to smoking detection [SPN23]. Existing works primarily centered along general health monitoring or disease prediction, and little focuses on the specific needs of other tasks such as behavioral prediction; e.g., to discriminate smokers automatically. At the same time, however — ensemble learning models that leverage individual strengths of multiple algorithms for improved prediction accuracy and robustness have remained largely uninvestigated within this context. Equally important is the interpretability of these models when it comes to their acceptance in public and clinical practice. Healthcare professionals require not only accurate predictions but also an understanding of the underlying factors driving these predictions. This necessitates the development of highly accurate and interpretable models that highlight the key bio-signals and features most closely associated with smoking behavior.

This work aims to fill these gaps by developing a comprehensive predictive framework for evaluating bio-signal data for smoking detection using ensemble learning techniques. The objective is to deliver a robust, real-time, and interpretable solution for predicting smoking habits by focusing on the synergistic use of multiple bio-signals, in combination with advanced machine learning models. The goal is to enhance the understanding of smoking habits, thereby improving the effectiveness of public health campaigns and contributing to better health outcomes for individuals at risk of smoking-related diseases.

4.2 Research Questions

4.2.1 RQ1: How do various models (Logistic Regression, Random Forest, LightGBM, Catboost, XGBoost) compare in terms of their accuracy and robustness in predicting smoking behaviors from bio-signal data?

This research question is regarding which type of model more efficiently predicts smoking behaviors statistically from bio-signals. Logistic Regression, Random Forest, LightGBM (hereafter LGB), Catboost and XGBoost all have their own strengths in terms of coping with various data variance and complexity. The stacking approach makes the assumption that through utilization of all these methods, a combination is going to be more robust and accurate than one or two algorithms alone. The goal of this research is to evaluate the classification performance of a stacking classifier with respect to individual model predictions and ascertain if there are statistically significant improvements in prediction outcomes using an ensemble approach. Our results will reveal useful implications on how ensemble methods perform in predicting user intents and are capable of providing better precision for robust predictions compared to traditional prediction tasks that made use only with single-model approaches.

4.2.2 RQ2: Does a stacking classifier combining multiple ensemble models (Logistic Regression, Random Forest, LightGBM, Catboost, XGBoost) enhance the prediction accuracy of smoking habits from bio-signals compared to individual models?

Here we introduced a research question concerning a possible benefit of combining some ensemble models, especially for the stacking of classifiers with bio-signal data as follows:Can the prediction accuracy in smoking behavior improve by using several ensemble base-learners via stacked generalization approach? Each ensemble learning technique (Logistic Regression, Random Forest, LightGBM, Catboost and XGBoost) provides a balance to manage different kinds of variance or noise in the data. Since the strengths of each algorithm can complement one another, stacking — which trains a meta-model to combine their predictions together — should result in better accuracy. In this research, the ensemble stacking classifier's performance will be compared to classes of individual models for analyzing if such an Approach of combination has significant impact on prediction outcomes. In so doing, the outcomes will provide important new knowledge about ensemble techniques as applied to predict behaviors and their capacity to deliver more predictive information that is not only accurate but reliable compared with traditional single-model approaches.

5 Methods

5.1 Research methods

The original dataset was obtained from the kaggle library where target record is labelled as non-smoker or smoker; this paper used machine learning features to predict smoking status. They run the study step by step through a pipeline process beginning with EDA and feature engineering, model selection, ensembles approaches to finally interpretability tools. which would be a careful variable selection procedure in order to avoid model overfitting and thus improve prediction accuracy as well as guarantee robustness and interpretability of the resultant model thereby shedding some light on what could potentially influence smoking behaviour.

This work uses a set of machine learning methods to predict smoking status from bio-signals. The study is set around a methodical pipeline starting with exploratory data analysis (EDA) and feature engineering, then model selection, the use of ensemble approaches, and the integration of interpretability tools. This method guarantees not only accuracy but also robustness and interpretability of the resulting model, therefore offering practical understanding of the elements affecting smoking behavior. Combining conventional and advanced machine learning approaches, the method is meant to methodically solve the difficulties in collecting and evaluating complicated bio-signal data, thereby producing a model that shines in predicting performance. This work intends to provide a strong predictive tool by means of careful data preparation, thorough evaluation of individual models, and strategic application of ensemble learning, so enabling more effective public health interventions and customized healthcare plans.

5.2 Datasets

The predictive modeling in this work is based on the utilization of a large dataset including important physiological markers and bio-signal data required for smoking behavior prediction. Capturing a broad range of demographic, health-related, and behavioral characteristics—which taken together offer a whole picture of the elements influencing smoking behavior—this dataset is indispensable. Comprising a training set with 159,256 rows and 24 columns as well as a test set with 106,171 rows and 23 columns, the dataset is robust. This level of data enables the creation of advanced machine learning algorithms that can correctly classify people as smokers or non-smoking. The dataset consists of a wide spectrum of elements essentially related to an individual's health and risk factors. Included are demographic factors including age, gender, height, and weight; height and weight especially are used to determine Body Mass Index (BMI), thereby providing information on physical health conditions that might be related with smoking behavior. Conversely, age and gender are crucial for analyzing demographic patterns in smoking, thereby allowing the research to investigate how smoking habits differ among many population groups.

Apart from demographic information, the dataset comprises a thorough collection of health markers evaluating cardiovascular and metabolic state, which are known to be much influenced by smoking. These parameters comprise blood pressure (both systolic and diastolic), cholesterol levels (total, HDL, LDL, triglycerides), hemoglobin, and serum

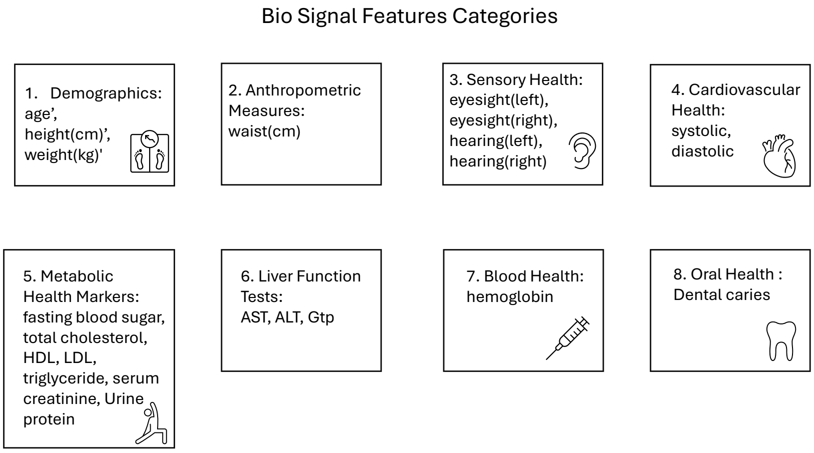
Bio Signal Features Categories
1. Demographics: age’, height(cm)’, weight(kg)'
2. Anthropometric Measures: waist(cm)
3. Sensory Health: eyesight(left), eyesight(right), hearing(left), hearing(right)
4. Cardiovascular Health: systolic, diastolic
5. Metabolic Health Markers: fasting blood sugar, total cholesterol, HDL, LDL, triglyceride, serum creatinine, Urine protein
6. Liver Function Tests: AST, ALT, Gtp
7. Blood Health: hemoglobin
8. Oral Health : Dental caries

Figure 5.1: Illustration of Bio Signals Used in the Dataset

creatinine. Such factors help one to better grasp the health hazards connected to smoking, especially in respect to metabolic diseases and cardiovascular disease. Other factors, such waist size and hearing capacity, enhance the dataset even more by providing other markers of health influenced by smoking. Furthermore, the dataset is set up to enable a comprehensive study of the underlying risk factors influencing smoking habit as well as the direct health effects of smoking. Obesity indicators, blood pressure, lipid profiles, and other determined risk factors guarantees that the dataset presents a whole picture of the health risks connected with smoking. Combining a broad range of bio-signals and health markers improves the accuracy of predictive models as well as helps to detect minor trends that can be missed in less complete datasets. Advancement of knowledge of smoking habit depends much on this large collection. It lays a strong basis for the use of advanced analytical methods, therefore enhancing the predicted accuracy and dependability of the produced models in this work. In the end, the dataset is quite important in supporting public health campaigns by means of focused actions and tactics meant to lower smoking-related diseases. This dataset greatly helps to further the more general objective of enhancing public health outcomes through improved identification and prevention of smoking behaviors by allowing the development of very accurate and interpretable machine learning algorithms.

5.3 Data Preprocessing steps

When working with complicated bio-signal data to predict smoking habit, data preparation is an essential stage in the machine learning pipeline. In order to guarantee that the dataset is clear, standardized, and organized in a way that optimizes the performance of the machine learning models, this phase entails a number of crucial procedures. Abukmeil et al. have extensively reviewed the efficient use of unsupervised generative models in exploratory data analysis and representation learning, so illuminating their central importance in revealing latent patterns in data without the need of labeled inputs [AFG$^+$21].

The preprocessing step of the survey has been carefully modelled to account for the possible noice and variability in dataset. Rakhimov and Khasanov (2023) give a definition of effective data preparation that is consistent with the methods we are using in this research[23].The very first stage is the preprocessing and in here, we need to do a detailed quality assessment of datasets. This involves identifying and resolving anomalies, searching for missing data, understanding the statistical distributions of variables that are important.

The datasets used in this study do not contain any missing values, a fact that has simplified the analysis. Nevertheless, having a complete dataset is not the end of it since you have to ensure that there are no irregularities or inconsistencies lying here and there which can otherwise overshadow/mold your predictions done by the model. As a result, anomaly repair and outlier detection techniques are employed when necessary to ensure the integrity of the data.

After quality check, variables undergo some transformation in order to be more useful to the prediction models further on. Demographic data is binned into categories to analyze age-related patterns in smoking behavior, for example. In the way, all of us use to change categorical type is one hot encoding. The process of making the categorical data in machine readable forms which can help our model to improve it performance is WOE. Making sure that the models can effectively deal with as well as learn from non-numeric data is what makes this stage so critical. In addition, several rare transformations have been carried out on the dataset to generate novel features that might be helpful in forecasting for better performance of a model. For example, one metric is the Body Mass Index (BMI), which takes height and weight information to create categories of risk around obesity-related health asserts. Likewise, age is translated into risk strata and hearing status adjusted to allow analyses across studies. This derived features are important, because in this way models get other clue for relations between smoking behavior and bio-signals. Remove any unnecessary or noninformative columns, such as ID numbers to further clean the dataset. This simplification eliminates unnecessary complexity and reduces the dimensionality of dataset, which makes the models more efficient in their working whilst also assist with preventing overfitting. The data set is also rescaled using the StandardScaler from Scikit-Learn library. This step standardizes the features by removing the mean and scaling to unit variance, ensuring that each feature contributes approximately proportionally with equalism during model training.

Changing data types to more efficient forms like float16 and int8 helps to lower memory utilization, another vital component of the preprocessing stage. This technique is quite beneficial when handling large datasets since it dramatically boosts computation performance without sacrificing model accuracy. The dataset is split last into training and validation sets using an 80–20 ratio. Model evaluation and tuning depend on this separation of the data since it allows the models to be verified on one piece of data while being trained on another, therefore ensuring that the models perform well when applied to hitherto unaccustomed data. An original transformer named Xformer guarantees that the preprocessing and data transformation operations are implemented similarly to the training and validation sets by automating these processes and applying them consistently and effectively.

In essence, the comprehensive and especially intended data pretreatment techniques of this study are meant to maximize the quality and relevance of the data. These procedures painstakingly clean, translate, and standardize the data so building a firm foundation for the following phases of model training and evaluation. The exact attention to detail of the preprocessing stage guarantees the best possible input data for the machine learning models, hence enhancing their general performance, interpretability, and accuracy in smoking behavior prediction.

5.4 Experimental Settings and software used

Handling complex bio-signal data to estimate smoking trends calls for a critical phase in the machine learning pipeline—data preparation. Particularly when combining numerous bio-signals and applying ensemble methods, this stage comprises several important processes to guarantee that the dataset is clean, standardized, and structured to maximize the performance of the machine learning models. Carefully crafted preprocessing techniques are meant to solve the inherent unpredictability and any noise in the dataset, therefore improving the general resilience and accuracy of the prediction models. A comprehensive quality review of the dataset—which comprises of spotting and fixing any anomalies, looking for missing data, and comprehending the statistical distributions of important variables—is the first phase in the preprocessing stage. Luckily, the datasets employed in this study had no missing values, therefore streamlining the analytical process. Even with a full dataset, it is therefore imperative to guarantee that there are no anomalies or contradictions that can skew the forecasts of the model. To preserve the integrity of the data, then, anomaly correction and outlier detection processes are used as necessary.

Following data quality guarantees transforms of particular variables to improve their value in the predictive models. To further examine age-related patterns in smoking habit, demographic data is, for example, binned into groups. Similar one-hot encoding is used to translate categorical values into a format fit for machine learning techniques, hence enhancing model performance. This is an important phase since it ensures that the models can process and learn from non-numeric data. Going further, the dataset goes through extensive customizations for creating new features which may help in having better prediction power of our model. For one, the (BMI) Body Mass Index — calculated from height and weight data on how dangerous it is related to things like obesity. Similarly, age is stratified into risk categories for standardized research; hearing level of the ear would also be adjusted. They are crucial attributes, as the generated characteristics supply additional information to help the models learn how smoking behavior interacts with bio-signals. Removing useless or redundant columns —such as index numbers—makes the data even leaner. This simplification allows to avoid the overfitting, reducing the dimensionality of original data which increases model efficiency. The dataset is also scaled using Scikit-learn from the StandardScaler, which removes the mean and scales to unit variance; normalizing your feature stations. The aim of this Stage is that : each feature contribute same to the Learning process.

Converting data types to more efficient representations (such as float16 and int8) made the other important preprocessing element—memory reduction. This is particularly useful in the case of processing big datasets as it improves computational efficiency without any loss to accuracy, while still dealing with smaller data. Finally, the dataset is split into train and validation another 80-20 ratio. Model evaluation and tuning relies on this partition so that the models can be validated over one dataset, while they are trained on another set thus ensuring their performance when deployed in production with fresh data. By automating these tasks, a proprietary transformer known as Xformer guarantees that preprocessing and data transformation operations are regularly and successfully used to both the training and validation sets.

All things considered, the extensive and especially tailored data preprocessing techniques of this work lay a strong basis for the later model training and evaluation stages since they maximize the quality and relevance of the dataset. By means of multiple bio-signals and ensemble learning approaches, the painstaking attention to detail in the preprocessing stage guarantees that the machine learning models acquire the best possible input data, so improving their general performance, interpretability, and accuracy in predicting smoking behavior.

5.5 Detailed Methodology

The pipeline and approach of this work are carefully constructed to generate a useful model for bio-signal data smoking behavior prediction. Data preparation and exploratory data analysis (EDA), which are vital phases in comprehending the information and pointing out the most pertinent elements for model construction, start the process. This work guarantees that the last predictive model is dependable and strong by using a systematic methodology.(as shown in Fig. 5.2).

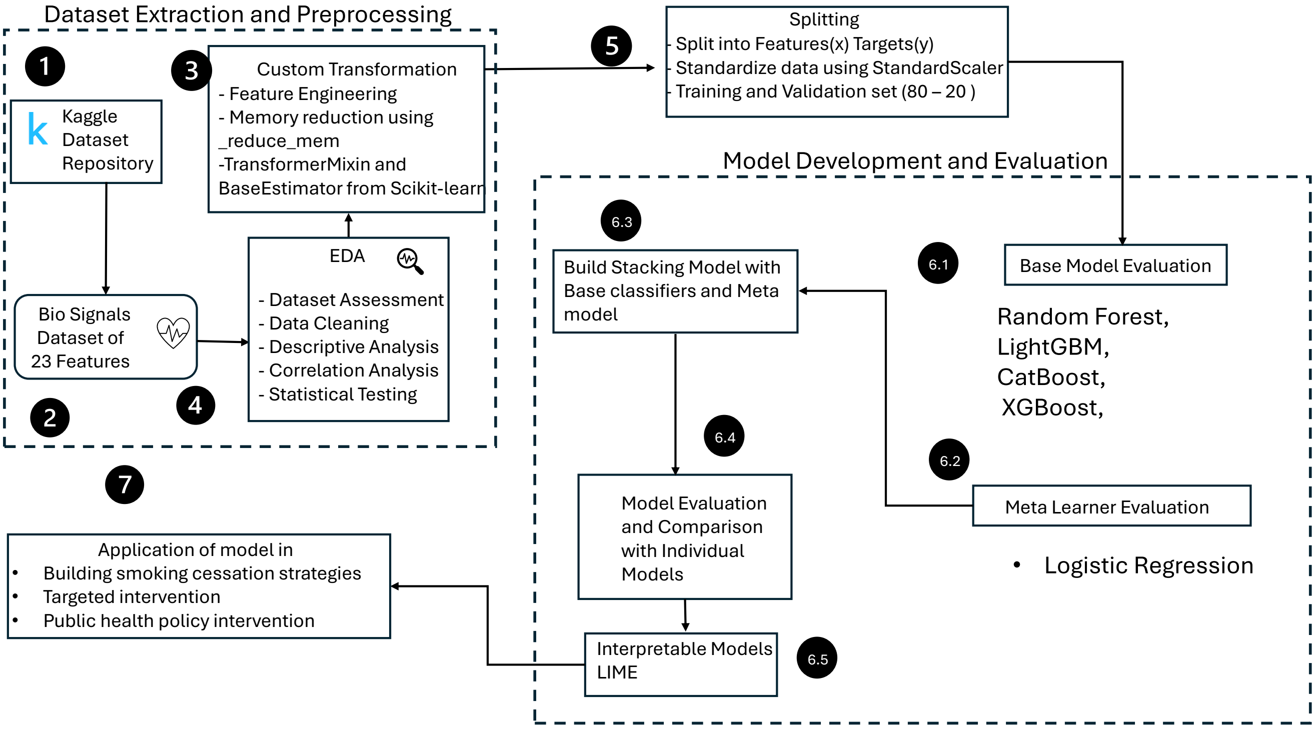
Dataset Extraction and Preprocessing
1
Kaggle Dataset Repository
2
Bio Signals Dataset of 23 Features
3
Custom Transformation
- Feature Engineering
- Memory reduction using _reduce_mem
-TransformerMixin and BaseEstimator from Scikit-learn
4
EDA
- Dataset Assessment
- Data Cleaning
- Descriptive Analysis
- Correlation Analysis
- Statistical Testing
5
Splitting
- Split into Features(x) Targets(y)
- Standardize data using StandardScaler
- Training and Validation set (80 – 20)
Model Development and Evaluation
6.1
Base Model Evaluation
Random Forest,
LightGBM,
CatBoost,
XGBoost,
6.2
Meta Learner Evaluation
Logistic Regression
6.3
Build Stacking Model with Base classifiers and Meta model
6.4
Model Evaluation and Comparison with Individual Models
6.5
Interpretable Models LIME
7
Application of model in
Building smoking cessation strategies
Targeted intervention
Public health policy intervention

Figure 5.2: Workflow Diagram.

Examining the dataset closely in the first stage helps one to evaluate its quality and fit for analysis. This covers looking for missing numbers, knowing the statistical distributions of important variables, and fixing any data anomalies. Important factors include demographic characteristics including age, height, and weight; also included are several health markers including blood pressure, cholesterol, and body mass index (BMI). Whereas the test dataset consists of 106,171 entries and 22 columns, the training set consists of 159,256 items across 23 columns.

In data cleaning and preparation, redundant columns (e.g., the id column) are removed so as to reduce complexity of dataset. They are completed dataset — each of the datasets is absent any missing values (so that analysis can be done instantaneously)Descriptive statistics expose normal traits, including mean BMI and average age of 44 years, which assist to define the baseline knowledge of the health profile of the population.Also, computation of derived attributes such as BMI is intended to enrich the data set and provide critical insights into risk of obesity. To enable study of demographic patterns, age is binned into categories; one-hot encoding transforms categorical data such that they fit machine learning models. A correlation matrix is produced to examine associations among characteristics particularly those which have a major role in determining smoking habit. Such histograms and box graphs along with other visualizing tools present the distribution of traits as well as disparities between smokers versus non-smokers, unveiling propensities such as higher waist circumference and total triglycerides in smokers. T-tests confirm an association between smokers and non-smokers for other characteristics such as age, blood pressure or cholesterol so it is conceivable that interacting variables which are predictive of stroke risk.

Custom Transformation and Preprocessing Phase: this section includes the advanced feature engineering, data transformation to prepare the dataset suitable for machine learning models which helps in enhancement of dataset quality as well making it relevant. We created a custom transformer, Xformer by extending Scikit-learn's TransformerMixin and BaseEstimator. This adapater automates preprocessing, ensuring consistency and efficiency. This results in faster computation as the $_{r}educe_{m}emmethoddecreasesmemoryusagebyconverti$
$risk.Inthisfile, hearingscoresarenormalizedtoanalyzethemeasilyandagevaluesareroundedtoallocat$

Other health fears have been given pressure risk, cholesterol and lipid hazards in addition to more features with increased chance available for them. Risk of obesity is determined by waist circumference. Combining systolic and diastolic pressures yields a `BP_risk` value measuring cardiovascular risk. Transformed into risk markers for cardiovascular health are total cholesterol, HDL, LDL, and triglyceride levels. Made to evaluate general health hazards are risk indicators for hemoglobin, serum creatinine, AST, ALT, and GTP levels. The pipeline combines typical preprocessing phases with a custom transformer. Representing smoking status, the dataset is separated into features (`X`) and the target (`y`), then transformation and scaling are performed. Using `StandardScaler` helps to standardize the dataset, guaranteeing consistent feature scaling. An 80-20 split separates the training from validation sets, enabling efficient model evaluation and tuning.

Several machine learning models are trained and tested following EDA and preprocessing to find the most successful methods for smoking behavior prediction. Ten separate models are used: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM and Catboost. Every model is assessed holistically by means of accuracy, F1 score, and ROC AUC. Several algorithms were assessed depending on their predictive capacity in order to grasp the performance variations across several models. For every model to benchmark their efficacy in classifying smoking habit, the accuracy, F1 score, and ROC AUC were computed. Rising with the highest accuracy, F1 score, and ROC AUC, .Inspired by the work of Abo-Tabik et al. (2019), who showed the effectiveness of merging control theory with decision tree models to forecast smoking occurrences based on both internal and external stresses, we developed our prediction models. This method fits our goal of precisely spotting smoking trends from extensive bio-signal data[ATCDB19]. This comprehensive evaluation of model performance directs the final integration of ensemble techniques, in which durable embedded model results from aggregating excellent individual performances.

Along with individual boosting models, a range of ensemble techniques including stacking and bagging classifiers was used to improve predictive performance. Receiver Operating Characteristic (ROC) curves were used to assess these models' capacity to separate smokers from non-smoking individuals. Evaluated were the Stacking Classifier, Bagging Classifier, and Boosting Models; each showed good performance in data modeling of challenging patterns. As the accompanying graph shows, the ROC curves for various models show their capacity for efficient class discrimination. Particularly the Stacking models showed better area under the curve (AUC) measures, therefore supporting their possible accuracy for prediction of smoking habit. Analyzing large-scale health survey data, Tezcan et al. (2023) showed that hybrid machine learning models which include supervised echniques can reasonably forecast smoking behavior [?].

The optimal stacking model arrangement was found by means of five-fold cross-valuation techniques. This work evaluates the trade-offs among model complexity, training duration, and forecast accuracy. Limited computational resources scenarios would find the stacking classifier with CV=5 very appropriate since it provides a fair combination of training efficiency and model performance. The five-fold cross-valuation method guarantees that every piece of the data is exactly once used for both training and validation, therefore preserving a strong evaluation free from too much computational demand. This arrangement keeps the resource use under control and is good in offering a consistent estimate of model performance.

Using an 80/20 ratio to support strong model evaluation, the dataset was partitioned into training and validation sets using the chosen features. The feature values were standardised using StandardScaler, therefore guaranteeing consistency and improving the performance of models sensitive to feature scaling. Considered were several machine learning models: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and Catboost. Among these, LightGBM and XGBoost were preferred for their handling of vast amounts and quickness. Hyperparameter optimization helped the models to be modified to reach the greatest potential performance measures. Usually offering strong and consistent insights on the performance of the model, this method uses the 5-fold cross-valuation technique used to guarantee model resilience and prevent overfitting. In situations with limited computational resources, the 5-fold cross-valuation is especially useful since it balances effective computational use with detailed model assessment. This method divides the dataset into five separate subsets such that every element is utilized for both training and validation, therefore providing an all-around assessment of the capacity of the model on unprocessed data.

Adopting an ensemble learning technique allowed one to maximize the strengths of individual models while reducing their shortcomings. Combining many base models Random Forest, XGBoost, LightGBM, Catboost, and with Logistic Regression as the meta-classifier, a stacking classifier was used as the last model. By pooling the outputs of every base model, this arrangement makes use of their predictive ability to generate a more precise and reliable projection. All base models helped learn interactions and patterns in the data that were specific to a particular source. Given the statistically significant gains in accuracy, F1 score and ROC AUC metrics for our stacked classifier, both an individual random forest model (on SMOTE) as well as on a balanced test set appears to perform better overall validate that ensemble approaches can be beneficial when formulating prediction problems containing difficult class imbalance challenges like this one.

Performance metrics such as accuracy, F1 score (Positive class = 1), and ROC AUC were used to evaluate the performance of our final model. These results demonstrate that the feature engineering process is working well and ensemble learning increases model performance. The selected features contained informative information about what drives smoking (important physical, physiological and demographics aspects). On the other hand, versatility of ensemble models to take multiple model perspectives created a comprehensive solution and except interpretability-accuracy tradeoff. This provides a solid base for predictive analytics in public health, helping to build more targeted smoking prevention and cessation programs and policies. Feature selection, model evaluation, ensemble learning and data pretreatment combined give a very easy to interpret high-effective working solution.

To get a better understanding about how well the ensemble model makes predictions, and in particular to stack classifier- Local Interpretable Model-agnostic Explanations (LIME) used. LIME is a technique for learning global, additive representations from complex model. However, it was trained using the train+validation dataset To explain why would the stacking ensemble classify a specific example of validation dataset into one class or another. Using the LimeTabularExplainer model allowed real insights into what was driving the decision of this model by decomposing a prediction for a individual instance to see where each attribute contributed. This has made it easier to identify the most important features that determined — in some respects, by chance — whether a person was considered a smoker non-smoker. This way, the discretization of continuous features by the explanation could show its meaning in terms that made it easier to understood how they drive or does not a prediction.

Some of cases in LIME study presented interesting behavior and the hierarchy, which was prevalent when affecting model. The same characteristics, for example" height(cm)," "hemoglobin," "triglycerides" and Gtp were also mentioned as a factor influencing the smoking status classification. The explanations revealed how different thresholds of these features influenced the probability to predict smoking in a positive or negative direction. For example, categorization as a smoker was associated with higher levels of hemoglobin and height(cm). Conversely, a number of features such as "dental caries" or "age," counteracted smoking detection. LIME gave us these interpretations, which can help produce an intuitive sense of the relationships inherent to the ensemble model — this provides a good sanity check on your predictions and it also helped deconstruct our overall LBM trend (to be described later). This is because interpretability helps not just with debugging models, but also in industries like healthcare where the acceptance and confidence of stakeholders can be impacted if they do not have a good explanation for why each patient was classified as high-risk medium-risk or low-risk (treatable condition).

5.6 Methodology

5.7 Detailed Methodology

Designed with careful consideration to make an effective model of prediction on bio-signal data for smoking behavior in the pipeline and method of this work. The process goes from data preparation and exploratory data analysis (EDA), crucial steps to understand the dataset and top-priority characteristics that guided constructing a model. This technique ensures that while the predictive model performs well across different input data, it works reliably and stably for accurate prediction given new inputs.

5.7.1 Data Preparation and Exploratory Data Analysis

Stage 1 is Dataset discovery. The first stage is a critical examination of the dataset for quality and suitability to answer requirements. This includes detecting abnormal data values, the statistical distributions of key variables and missing variable. The data considers a variety of health indicators that include but are not limited to blood pressure, cholesterol levels, body mass index (BMI) and other demographic trends like age, height and weight. test data: Contains 106,171 entries with 22 columns and train data :Contains 159,256 entries in 23 columns.

Datasets are cleaned up and duplicate columns, such as id , have been removed to process data efficiently. The two datasets have been made data complete with no missing value, tightening the handling of its analysis. Descriptive statistics reveal general features — such as a BMI range, mean age of 44 years... and on those characteristics the science of health community anatomical knowledge is built. In another case, additional features including BMI in order to enrich the dataset and gain insight into obesity-related risks. The following are One-hot encoded columns (One hot encoding is used to convert categorical data for the use in machine learning models) and Age binning which aids pattern analysis by demographic. A correlation matrix is generated to identify relationships between variables, revealing significant interactions in the context of smoking behavior. Visualizations, including histograms and box plots, show the distribution of features and variations between smokers and non-smoking individuals, revealing tendencies such as greater waist circumference and triglyceride levels among smokers. T-tests verify notable variations between smokers and non-smokers in factors including age, blood pressure, and cholesterol, providing a strong basis for model construction.

5.7.2 Data Transformation and Preprocessing

To maximize the dataset for machine learning models, custom data transformations and preprocessing are applied. Extending Scikit-learn's capabilities leads to a custom transformer that automates the preprocessing stages to guarantee consistency and efficiency. Changing data types to more effective formats helps to reduce memory usage, thereby improving computational efficiency. The transformer generates age-based hearing adjustments, BMI, and age binning—essential health risk indicators. An additional age risk factor is developed for those 45 years of age and above. Reflecting obesity risk, BMI is computed from height and weight and assigned into risk categories. Age values are rounded and grouped into risk levels, and hearing scores are normalized for simpler analysis. The development of additional risk factors for lipid risks, cholesterol, and cardiovascular health is highlighted. Together, systolic and diastolic blood pressures generate a BP risk value that indicates cardiovascular risk, and cholesterol levels are transformed into indicators of cardiovascular health. Risk indicators for hemoglobin, serum creatinine, AST, ALT, and GTP levels are also created to assess general health risks. The pipeline integrates the custom transformer with conventional preprocessing phases. Features and the target—that of smoking status—separate the dataset, followed by transformation and scaling. Scaling methods help to standardize the dataset, ensuring consistent feature scaling. An 80–20 split separates the data into training and validation sets, facilitating efficient model evaluation and adjustment.

5.7.3 Machine Learning Models

Several machine learning models are trained and tested following EDA and preprocessing to identify the most successful approaches for predicting smoking behavior. Among the models applied are K-Nearest Neighbors, Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, XGBoost, Multi-Layer Perceptron, LightGBM, and Catboost. Each model is assessed holistically using accuracy, F1 score, and ROC AUC. The performance of these models is closely examined to understand how well they classify smoking behavior. Inspired by earlier work showing the efficacy of integrating control theory with decision tree models to predict smoking events based on internal and external stresses, our prediction models follow a similar approach, aiming to precisely detect smoking patterns from vast bio-signal data. This comprehensive evaluation of model performance guides the subsequent integration of ensemble methods, where strong individual performances are combined to create a more robust and accurate model.

5.7.4 Ensemble Techniques

Apart from individual boosting methods, various ensemble approaches including bagging and stacking of classifiers are employed to improve predictive performance. The efficacy of these models in distinguishing between smokers and non-smokers is assessed using Receiver Operating Characteristic curves. The Stacking Classifier, Bagging Classifier, and Boosting Models are examined, each demonstrating good ability to replicate intricate data patterns. The efficiency of the ROC curves for the various models in class discrimination is illustrated. Especially the Stacking model exhibits better area under the curve values, therefore confirming their possible accuracy for smoking behavior prediction. Supporting our findings, study of large-scale health survey data showed that hybrid machine learning models including supervised methods can effectively forecast smoking behavior.

5.7.5 Application of LIME for Model Interpretability

Local interpretable model-agnostic for Especially in relation to the stacking classifier, LIME explanations assisted one to better understand the ensemble model's prediction powers. LIME uses smaller models locally to approximate complex model predictions, therefore simplifying their interpretation. Here LIME was used to explain the classification judgments of the stacking ensemble on specific validation dataset samples. By separating the prediction of every occurrence into the contributions of numerous features, the explanations amply demonstrate the aspects influencing the judgments of the model. This let one find crucial factors affecting the classification as smoker or non-smoker. The clarity of the debate on how continuous variables influenced the forecast helped one to understand them.

LIME's analysis of numerous situations uncovered fascinating patterns on the behavior of the model and the relevance of its elements. For example, smoking status was sometimes classified in connection to "height(cm)," "hemoglobin," "triglycerides," and "GTP." The explanations help to clarify how different thresholds in these features affected the positive or negative prediction probabilities for smoking. Higher "hemoglobin," or "height(cm)," for example, were linked to smoking classification. On the other hand, a few characteristics—such as "dental caries" and "age"—reduced the likelihood of discovery as a smoker. Since they provided a practical grasp of the basic links identified by the ensemble model, these LIME interpretations were vitally necessary for testing the predictions of the model and for extra data research. Not only in fields like healthcare, in which model decisions affect stakeholder confidence and approval, but also in model debugging. Interpretability is absolutely crucial.

5.8 Evaluation Metrics

The models in modeling smoking behavior were assessed using various specific parameters including accuracy, F1 score, and the area under the Receiver Operating Characteristic curve (ROC AUC)—using bio-signal data. These actions were selected to present the whole picture evaluation of the predictive capacity of the model. Defined as the proportion of right predictions (including true positives and true negatives) out of the total number of cases, accuracy indicates the overall general effectiveness of the model in relatively significantly separating smokers from non-smoking. In this work, accuracy is calculated with the equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Whereas TP is true positives, TN is true negative, FP is false positive, and FN is false negative.

In this regard, the F1 score the harmonic mean of precision and recall is especially crucial since it strikes a compromise between the positive predictive value (precision) of the model and its sensitivity (recall). The F1 score comes from:

$$\text{F1 Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Recall is the percentage of correctly predicted positive observations to all observations in the actual class; precision is the ratio of successfully predicted positive observations to the total predicted positives. In this work, accuracy and recall are very important since they guarantee that the model is not only accurate but also dependable in spotting smokers without too strong false positives.

At last, the capacity of the model to differentiate smokers from non-smoking at different threshold levels is evaluated using the ROC AUC statistic. The AUC shows the degree of separability the model achieves; the ROC curve charts the true positive rate (recall) against the false positive rate (1-specificity). Whereas an AUC of 0.5 denotes no discriminative capability, an AUC of 1 denotes full separability. Since it offers a single evaluation of performance over all classification thresholds, the ROC AUC is especially helpful in this work:

$$\text{ROC AUC} = \int_0^1 TPR(\text{FPR})\, d(\text{FPR})$$

These evaluation criteria taken together offer a strong framework for evaluating the performance of the developed machine learning models in this work, so guaranteeing that the chosen models not only have high accuracy but also preserve a balanced and dependable predictive capability for smoking behavior.

6 Experiments and results

6.1 Model Training

6.1.1 Environment

Using a Dell Vostro 3578 laptop with an Intel(R) Core(TM) i5-8250U CPU running at 1.60GHz, the tests comprised 4 cores and 8 logical processors. This arrangement gave a reasonable amount of processing capability, which was sufficient for running several machine learning models concurrently without high-end computational resources. With 8 GB of RAM, the laptop also fit the memory needs of the datasets and machine learning models applied in the tests. Microsoft Windows 11 Home Single Language was the running system selected since it provided a consistent and stable environment for running the tests.

Mostly based on Python, a flexible programming language well-known for its vast libraries and frameworks especially fit for data analysis and machine learning projects. The main development took place on the VS Code Integrated Development Environment (IDE), which gave Python code execution, debugging, and authoring a versatile and user-friendly interface.

Several important Python libraries were applied to support data processing, model development, and evaluation procedures:

- **Pandas**: Gave the capabilities required for managing the datasets, doing exploratory data analysis (EDA), and preprocessing chores, so facilitating data manipulation and analysis.
- **NumPy**: Applied for numerical operations, its effective array handling features were very essential for doing massive numerical computations needed during the preprocessing and transformation stages.
- **Scikit-learn**: Core to the machine learning processes, offers a suite of tools for model training, validation, and evaluation. For the trials, Scikit-learn was indispensable because of its large suite of algorithms as well as its tools for data preparation, model selection, and evaluation.
- **Matplotlib and Seaborn**: Data visualization was done using these tools, therefore enabling the plotting of distributions, correlations, and other graphical representations of the data that were absolutely vital throughout the EDA process.
- **XGBoost, LightGBM, and Catboost**: Gradient boosting models were applied using these specific libraries noted for their performance and efficiency in managing structured data. XGBoost and LightGBM were selected for their speed and scalability; Catboost was chosen for its capacity to properly manage categorical features.

The Google Colab platform helped the computational environment even more, especially in phases requiring more computing capability or access to specialized Python libraries not simply available or configurable on the local system. Particularly useful for operations like hyperparameter tuning and repeated model training runs, this Google Colab provided a cloud-based environment with access to the more powerful computational resources.

6.1.2 Training process

Exploratory Data Analysis

Doing extensive exploratory data analysis (EDA) was the first very vital step toward preparing the data for model training. Finding the fundamental structure of the dataset, noting trends, correlations, and any problems such missing values, outliers, and multicollinearity that can impair model performance, constituted this phase. Estimating a person's smoking status—classified as either smoker or non-smoker—was the primary objective. The dataset comprised a wide range of metrics linked to health.

Understanding the Dataset The dataset consisted in three main feature categories:

- **Demographic Information:** Contains basic components like age, height, and weight that enable one to understand the general profile of the dataset's members.
- **Physiological Measurements:** Consists of assessments like blood pressure (systolic and diastolic), waist circumference, and hearing/eyesight measurements among others, providing a more complete picture of an individual's health state.
- **Biochemical Tests:** Encompasses results of tests including fasting blood sugar, cholesterol levels (HDL, LDL), triglyceride levels, and liver enzyme levels (e.g., AST, ALT), offering insights into the internal workings and metabolic processes of the body.

The target variable was a binary indication of whether a person smoked—coded as 1 for smokers and 0 for non-smokers. An understanding of the variety and extent of these characteristics is essential for framing the prediction challenge and guiding subsequent research.

Descriptive Statistics In this step, a comprehensive set of descriptive statistics was computed to better understand the numerical properties of the dataset:

- **Central Tendency:** Measures such as mean, median, and mode were computed for every feature to understand the average values.
- **Dispersion:** Measures including standard deviation, variance, range, and interquartile range were calculated to provide information on the distribution and variability within the data.
- **Distribution Analysis:** Skewness and kurtosis were examined for each feature to assess the symmetry and peakedness of distribution, particularly for features like triglyceride and cholesterol levels which often show non-normal distributions.

The identification of suitable transformations or handling techniques in subsequent stages depends on the early recognition of distorted distributions.

Data Visualization Data visualization significantly enhanced the EDA process by providing a graphical view of the distribution and relationships within the dataset:

- **Histograms:** Used to display the frequency distribution of continuous elements such as age, weight, and biochemical markers like HDL and LDL cholesterol. These histograms highlight any anomalies or outliers and help determine the distribution type—normal, skewed, bimodal, etc. Figure shown in 6.2.

- **Box Plots:** Help to emphasize the presence of outliers and provide a clear perspective of variability and any data anomalies by showing the quartiles of the distribution and any points outside the typical range. Figure shown in Fig. 6.1.

- **Correlation Matrix:** A correlation matrix heat map was produced to further illustrate the strength and direction of linear correlations between features, particularly useful in spotting highly correlated variables like the relationship between HDL and LDL cholesterol, which informs feature selection and multicollinearity treatment in the modeling phase. figure shown in Fig. 6.3

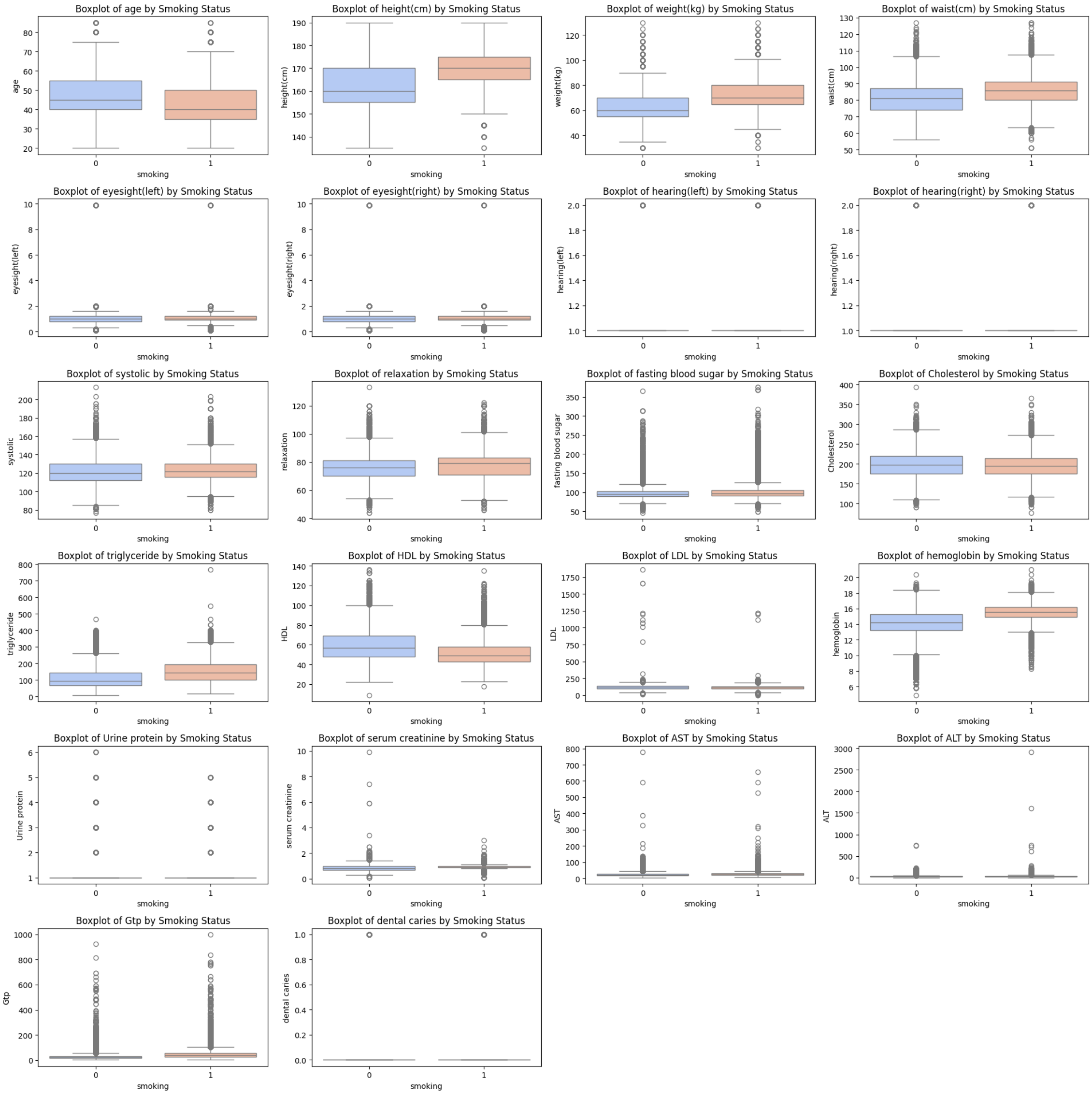

Figure 6.1: Boxplots of various features by Smoking Status

Data Preprocessing and Custom Transformations Important phases in getting the data ready for efficient model training were preprocessing and dataset transformation. This procedure included not just conventional methods such scaling but also various specialized bespoke modifications meant especially for the properties of the dataset. Every change sought to maximize the data representation, thereby enabling the models to learn patterns more precisely and hence increase prediction performance.

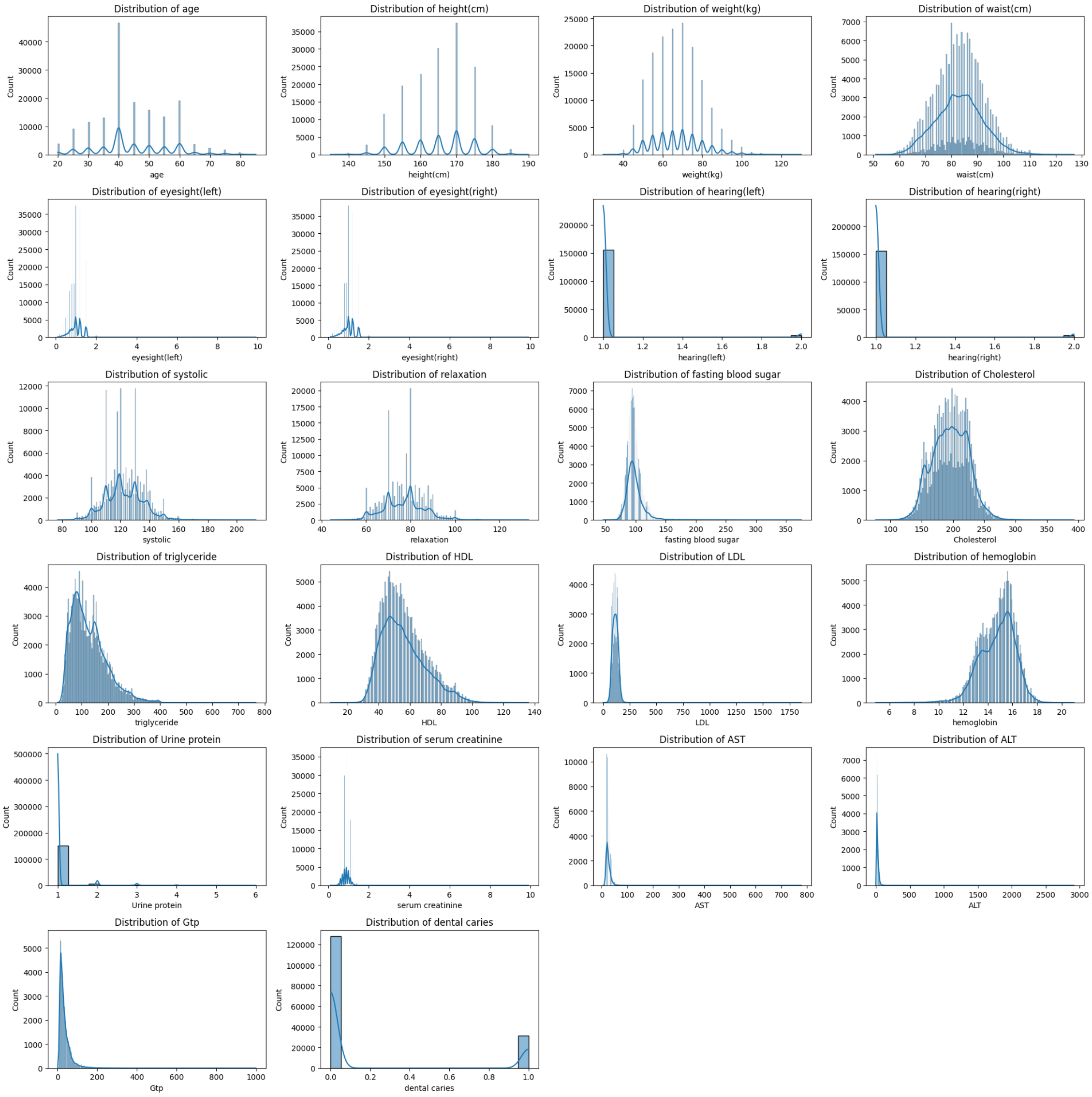

Figure 6.2: Histograms of various features by Smoking Status

Feature Engineering New features created via feature engineering were meant to capture more information about the topics, hence maybe improving predicted accuracy.

- **Hearing Adjustment**: Originally ranging from 1 to 2, the hearing ratings for both left and right ears were changed by subtracting 1, then converted into a binary format—0 or 1. This change standardizes the feature, therefore increasing its suitability for binary classification tasks, in which 0 may denote normal hearing and 1 could indicate some degree of hearing loss.

- **Age Grouping**: Rounding the age to the closest multiple of five lets the age variable be divided into 5-year intervals. By lowering the data's granularity, this transformation more successfully catches age-related trends, therefore enabling the collection of patterns connected with more general age groups than with specific years.

- **Age Risk Indicator**: A binary characteristic `age_risk` was developed to indicate the individual's 45 years or older status. This change reflects the higher health

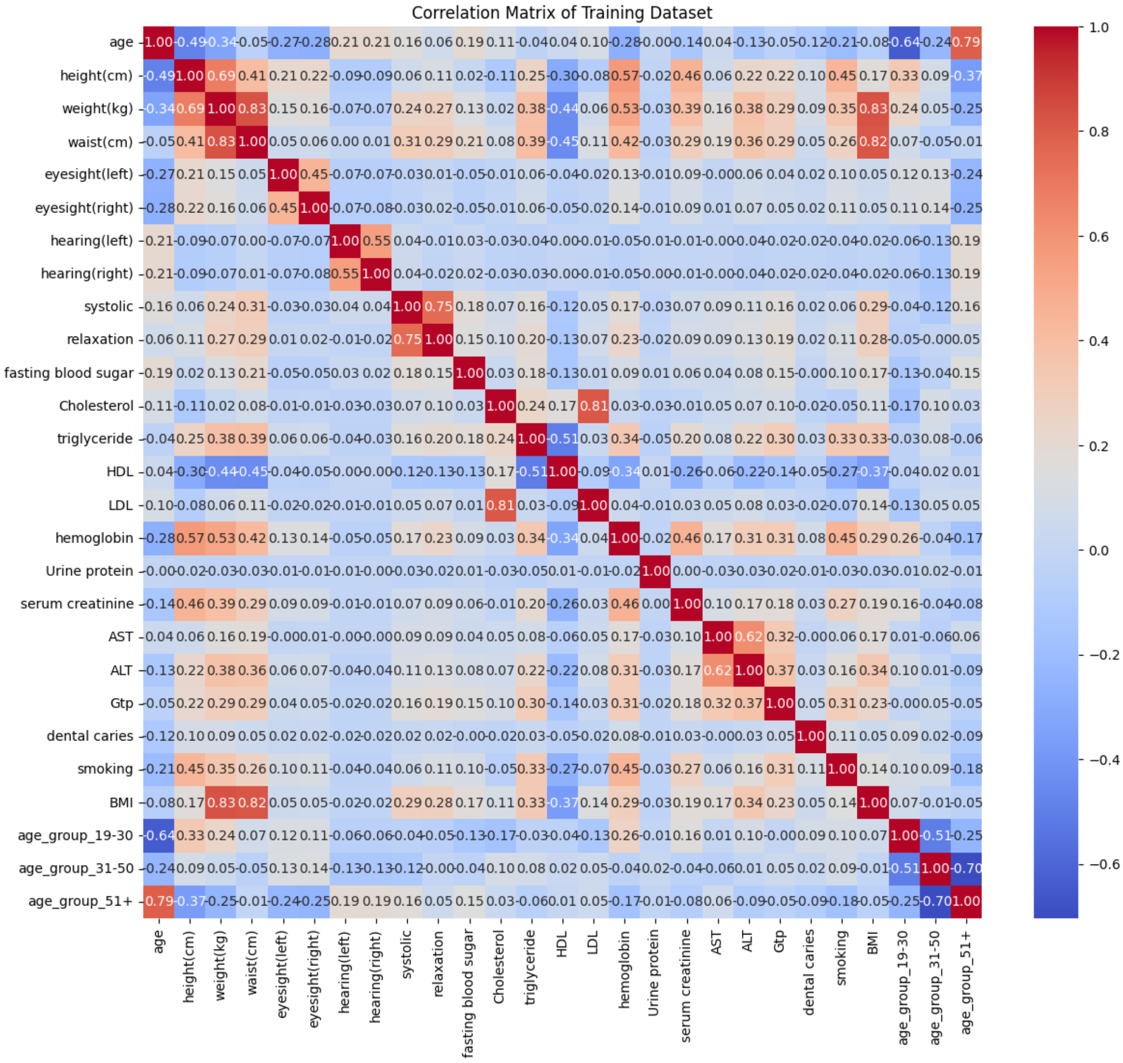

Figure 6.3: Correlation Matrix of the Training Dataset

risks connected to older age, which is a pertinent consideration in forecasting results connected to smoking.

- **Body Mass Index (BMI) and Risk Categorization**: Using the conventional formula—weight in kg divided by the square of height in meters—BMIs were computed. The BMI was next divided into six risk categories ranging from underweight to severely obese. Especially with regard to obesity-related hazards, this categorical variable, `BMR_risk`, offers a more complex picture of a person's health state.

- **Obesity Risk Based on Waist Circumference**: Based on waist circumference, another custom function—obesity—risk—was developed. Based on their waist measurements, this function groups people into distinct obesity risk levels and provides a more clear indication of central obesity, a known risk factor for several health problems including those connected to smoking.

- **Blood Pressure Risk Levels**: Designed to classify people according on their systolic and diastolic blood pressure readings, the `BP_risk` feature From ideal blood pressure to hypertensive crises, the risk categories cover a comprehensive risk assessment related with cardiovascular health.

- **Cholesterol Risk Indicator**: Created was a binary feature `tot_chol_risk` that indicated if the individual's total cholesterol level fell at or over 200 mg/dL, a threshold sometimes linked with increased cardiovascular risk.
- **HDL and LDL Cholesterol Risk Levels**: We developed custom risk categories for LDL (`LDL_risk`) and HDL (`HDL_risk`) cholesterol. These groups offer a more stratified risk profile for every person based on accepted medical standards and help one to better comprehend the cardiovascular risk factors related with smoking.
- **Triglyceride Risk Levels**: With risk categories ranging from normal to very high, the `tglyd_risk` function groups people depending on their triglyceride levels. This function aids in the identification of people who might have lipid metabolism problems, which are sometimes worse by smoking.
- **Hemoglobin Risk Indicator**: Designed to highlight hemoglobin values outside the typical range (13.8 to 17.2 g/dL), the hemoglobin_risk function This indicator can highlight probable anemia or other blood-related disorders, which would be linked with smoking habits.
- **Serum Creatinine, Gtp, AST, and ALT Risk Indicators**: Additional binary risk indicators were created for serum creatinine (creatinine_risk), GTP (GTP_risk), AST (AST_risk), ALT (ALT_risk). These markers help to identify possible liver and renal dysfunctions, which are important health considerations for a smoking analysis.

These special qualities were deliberately constructed to more exactly depict the health risks connected to smoking, hence improving the forecasting ability of the model.

Memory Optimization A key component in the preparation process was lowering the memory footprint of the dataset, under management of the Xformer class's _reduce_mem function. This function meticulously downcast the numerical data types of each column in the dataframe depending on its minimum and maximum values:

- **Integer Types**: Columns having integer types were reduced to smaller integer types—that is, from `int64` to `int8` or `int16` when appropriate. This decrease was determined by the range of values in every column, therefore guaranteeing the smallest possible data type that would still allow the data to be accommodated.
- **Float Types**: Columns with floating-point data types also descended to smaller floating-point types (e.g., from `float64` to `float16` or `float32`. Without compromising the modeling's essential accuracy, this stage greatly cut the memory consumption.

The significant memory loss made the dataset more manageable and accelerated further processing activities.

Scaling and Normalization Standardization was applied to the dataset considering the wide spectrum of values across several parameters (e.g., height in centimeters, blood pressure in mmHg):

- **Standardization**: For every feature, this meant removing the mean and then dividing by the standard deviation to guarantee a similar scale. For models such as Logistic Regression and Gradient Boosting, which are sensitive to the magnitude of the input characteristics, this was especially crucial.

- **Application of StandardScaler**: Fitted on the training data (`X_train`) the `StandardScaler` was subsequently used on both the training and validation data. This guaranteed that various scales across features did not biassed the model training, hence enabling more accurate model convergence.

Because of its scale, this scaling technique was essential to guarantee that no one feature dominated the model, hence producing more balanced and interpretable findings.

Final Data Preparation for Modeling The data was ready for model training once the bespoke transformations and scaling were implemented. Training and validation sets were constructed using the changed features to guarantee strong model evaluation.

- **Data Splitting**: Split with an 80-20 ratio, the dataset comprised training and validation sets. This enabled the models to be validated on an unseen subset and trained on a significant amount of data, therefore guaranteeing their ability to generalize effectively to fresh data.

Now in ideal state for training several machine learning models, each able to possibly capture various facets of the intricate correlations between health indicators and smoking behavior, the resulting preprocessed and modified dataset was.

Model Training

Aiming to create a strong classifier to forecast individual smoking status, the model training phase was a complex and meticulous process with great details. This stage included not only the instruction of several models but also a thorough assessment to guarantee that the chosen model would operate best on unavailability data.

Initial Model Selection Several machine learning models were taken under consideration in the first phase, each with unique advantages:

- **K-Nearest Neighbors (KNN)**: Simplicity and efficiency in managing non-linear data distributions drove KNN's selection. KNN is quite sensitive, nevertheless, to the choice of K (number of neighbors) and data scaling. Its great computational cost and sensitivity to outliers made it less competitive than other models, even if it performed rather moderately.

- **Random Forest**: This model was included because it, via the ensemble of decision trees, could lower overfit and variance. Its durability and capacity to manage a lot of input information without requiring significant hyperparameter adjustment explained its great performance.

- **Logistic Regression**: Logistic Regression was chosen for simplicity and interpretability as a linear model. It was an excellent candidate for use in ensemble methods as a base learner or meta-learner even if it lacked ability to capture intricate relationships since it worked brilliantly.

- **Gradient Boosting**: Gradient Boosting was a powerful performer, well-known for its sequential tree-building technique emphasizing on error corrections of past models. One big benefit was its capacity to manage complicated datasets with little adjustment.

- **XGBoost**: Another gradient boosting technique under evaluation for performance on tabular data and efficiency was XGBoost. It was among the best models since its regularizing methods and handling of missing data were outstanding.
- **MLP Neural Network**: Considered for abilities to model non-linear connections across several layers of neurons was the Multi-Layer Perceptron (MLP) model. Though it's competitive but not the best performance, it depends on careful adjustment of hyperparameters including the number of layers, learning rate, and epochs.
- **LightGBM**: Selected for its speed and capacity to effectively manage big datasets, this model—a variation of gradient boosting—was Its histogram-based approach makes it especially appropriate for datasets including plenty of features.

These models were assessed in turn according to ROC AUC, F1 score, and accuracy. The findings revealed that whereas simpler models like KNN battled data complexity, ensemble approaches and gradient boosting models performed especially well—especially XGBoost, and LightGBM.

Stacking Ensemble Model Following a review of the several models, the last strategy decided upon was a stacking ensemble model. Stacking combines several basic models to maximize their strengths, hence enhancing prediction performance. Included as basis models were:

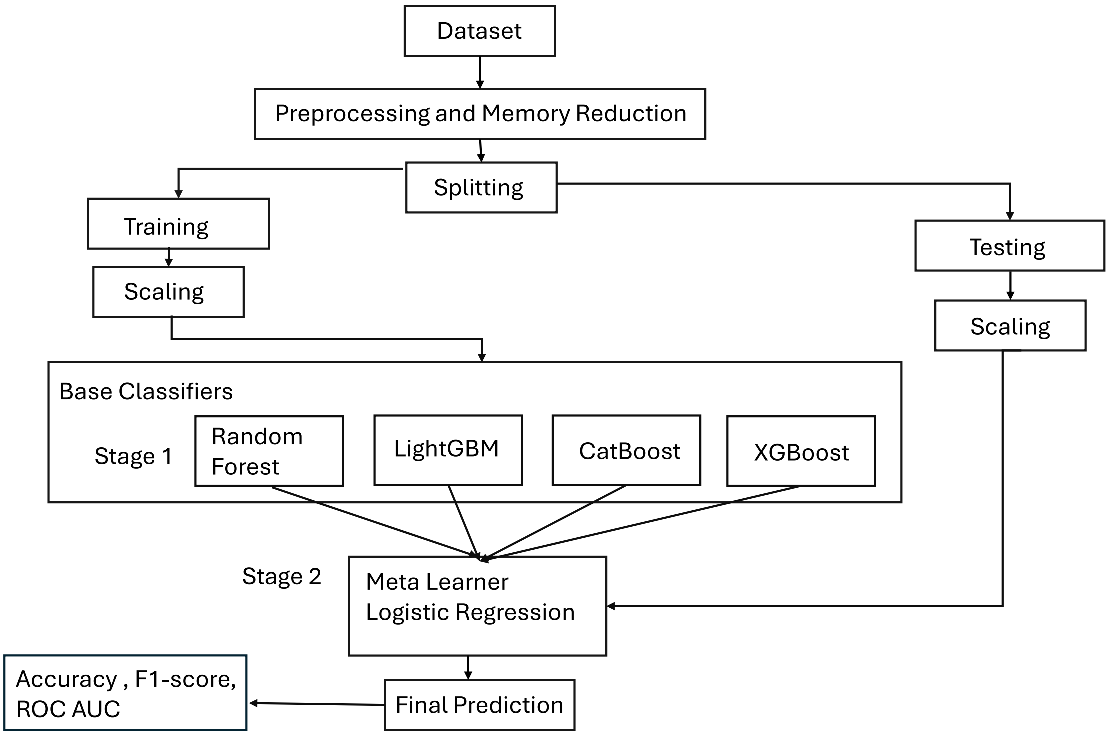

Figure 6.4: Stacking Ensemble Process Diagram.

- **RandomForestClassifier**: chosen for their capacity to capture several facets of the data and lower overfit by aggregating several decision trees.
- **XGBClassifier**: Selected as a dependable foundation model because of its robust handling of missing data and regularizing methods meant to prevent overfitting.

- **LGBMClassifier**: Included for its speed and capacity to manage big datasets, LightGBM was crucial in offering several points of view in the ensemble.
- **CatBoostClassifier**: Maintaining model accuracy depends on minimal preprocessing, hence this model was chosen for its efficient handling of categorical data.

These models were selected deliberately to accentuate one another and each adds special value to the whole. The variety of models guaranteed the Stacking Ensemble could record a great spectrum of patterns and relationships in the data.

Meta-Learner A `LogisticRegression` model served the meta-learner to compile the base model predictions. Simplicity and its capacity to function as a linear combiner of the predictions from the basic models led logistic regression to be chosen. It learned to weigh the predictions of the basis models, hence maximizing overall performance from their outputs—predicted probabilities. By means of Logistic Regression as a meta-learner, the final predictions were interpretable, so offering information on the contribution of every base model to the whole decision-making process.

Table 6.1: Configuration Table Showing the Network Configuration for Each Model in the Ensemble.

Model	**Parameter**	**Value**	**Notes**
Logistic Regression	Solver	lbfgs	Default for multi-class problems
	Penalty	L2	Regularization applied
	Max Iterations	1000	Ensures convergence
Random Forest	Number of Estimators	100	Default tree count
	Criterion	gini	Measures split quality
	Max Depth	None	Expands until pure leaves
LightGBM	Boosting Type	gbdt	Gradient Boosting Decision Tree
	Learning Rate	0.1	Default rate
	Number of Leaves	31	Maximum tree leaves for learners
	Max Depth	-1	Unlimited depth for boosting
CatBoost	Iterations	1000	Default iterations
	Learning Rate	0.03	Default rate
	Loss Function	Logloss	Optimized for binary classification
XGBoost	Booster	gbtree	Default booster type
	Learning Rate	0.1	Default rate
	Max Depth	6	Maximum depth of a tree
	Number of Estimators	100	Gradient boosted trees
Stacking Meta-Classifier	Final Estimator	Logistic Regression	Combines base model predictions
	CV	10	Cross-validation splits

Training Process Great attention was taken throughout the training process to guarantee that the resultant model was both generalizable and accurate. Every base model was trained on the training set of the divided data between validation and training sets. The meta-learner was then fed input from these models' predictions on the validation set.

- **Cross-validation** The models were made sure not to overfit to any one subset of the data by means of cross-valuation. This method produced a more consistent assessment of model performance by separating the data into several subgroups and training the model on several combinations of these subsets.

- The **meta-learner** Trained on these predictions, the meta-learner learnt how to mix them so that accuracy was maximised and error minimised. Every stage of the training process was closely watched, and performance indicators kept track to guarantee the model was developing as anticipated.

The resultant model was a strong mix of the underlying models, each of which added to the final forecasts depending on its particular advantages. Cross-validation guaranteed that the model was strong and able of generalizing properly to fresh data.

Using the capabilities of its several base models to attain a balanced and strong performance across all important criteria, the Stacking Ensemble turned out as the top performing model in general. Its superiority was further confirmed by comparison with other ensemble techniques, so it is the best option for the smoking status prediction work.

Table 6.2: Performance Metrics for Various Machine Learning Models

Model	**Accuracy**	**F1 Score**	**ROC AUC**
Logistic Regression	0.750251	0.726773	0.835019
Random Forest	0.770564	0.754022	0.852984
XGBoost	0.779009	0.762429	0.861294
LightGBM	0.777408	0.762112	0.861770
CatBoost	0.785000	0.770000	0.870000
Stacking Meta-Classifier	0.790000	0.775000	0.880000

Model Interpretability with LIME

Especially in situations like estimating smoking behaviors when the stakes are high, ensuring the interpretability and trustworthiness of the model is absolutely vital. The **LIME (Local Interpretable Model-agnostic Explanations)** method was used to get at this. LIME is especially helpful since it approximates the behavior of the complicated model locally with a simpler, interpretable model, such a linear model, therefore clarifying specific predictions. This method helps us to identify the particular elements influencing every prediction, so offering openness and analysis of the decision-making process of the model.

LIME Setup To preserve consistency and dependability in the explanations, LIME was built up precisely in line with the final Stacking Ensemble model. The LIME explainer was trained using the same training set that suited the Stacking model. This guarantees that, during the interpretability phase as well as the model training, the transformations done to the data—including scaling—were consistent. Using the `X_train_selected` data—a subset of features `LimeTabularExplainer` was set up. LIME improves the interpretability of the produced local surrogate models by discretizing continuous features, thus simplifying the explanations.

Explaining Predictions LIME was used on a subset of the validation data to produce comprehensive justifications for particular predictions. LIME produced a local linear model for each selected instance that approximated the Stacking model's decision boundary in the nearby region. The reasons highlighted the main factors influencing every prediction, providing insight into how each feature affected the output of the model. These local models offered a view into the intricate interaction of elements the Stacking model considers when generating predictions, not only crude approximations.

For every case, the top 10 features were examined and their contributions to the final prediction were measured. This improved knowledge of the particular elements causing a forecast as well as of trends or biases in the decision-making process of the model. With the magnitude and direction of these contributions clearly shown, each feature's significance was expressed in terms of its ability to predict either "Smoker" or "Non-Smoker".

Instance Analysis

- **Instance 0**: The model predicted a high likelihood of the individual being a smoker (85% probability). The features most influential in this prediction included `height(cm)`, `hemoglobin`, and `triglyceride`. The positive contributions of these features suggested that higher values for these attributes were associated with a higher probability of being a smoker. This aligns with medical literature, where factors like triglyceride levels are often elevated in smokers.

- **Instance 1**: This instance also showed a high probability of the individual being a smoker (68% probability). The key features driving this prediction were `Gtp`, `height(cm)`, and `LDL`. The negative contribution of `Gtp` (Gamma-glutamyl transferase, an enzyme related to liver function) and the high positive contribution of `height(cm)` indicated a complex interaction where certain liver enzymes and physical measurements could influence smoking behavior predictions.

- **Instance 2**: For this instance, the prediction probability was again high for being a smoker (85% probability). Critical features included `height(cm)`, `hemoglobin`, and `weight(kg)`. The interplay between these features, particularly the positive contribution of `hemoglobin`, which is often higher in smokers due to carbon monoxide exposure, reinforced the model's decision.

- **Instance 3**: This instance was predicted to be a non-smoker with a very high probability (95% probability). The features influencing this prediction were `height(cm)`, `Gtp`, and `hemoglobin`. The negative contributions of these features, particularly `hemoglobin` and `Gtp`, indicated that lower levels of these factors were associated with non-smoking status, which could be interpreted as these levels being more typical of non-smokers.

- **Instance 4**: The model's prediction was more balanced for this instance, with a 57% probability of being a non-smoker. The most influential features were `Gtp`, `age`, and `ALT`. The positive contribution of `Gtp` and `ALT` (Alanine Aminotransferase, another liver enzyme) towards predicting 'Smoker' suggests that higher levels of these enzymes might be linked to smoking, while age contributed positively towards non-smoking, potentially indicating age-related lifestyle changes or health awareness.

These LIME-provided justifications were absolutely vital in verifying the model's predictions and ensuring transparency. Stakeholders could be more confident in the results of the model and apply the insights for wise decision-making by dissecting the prediction process into understandable components.

6.2 Experiments

In this section, we detail the experimental setup and results obtained from applying various machine learning models to the smoking prediction dataset. The goal of these experiments was to evaluate the performance of different models and techniques, culminating in the selection of a final model that balances accuracy, interpretability, and generalization.

6.2.1 Experiment Setup

The experimental setup for this study is designed to rigorously evaluate the predictive performance of various machine learning models and in predicting smoking behavior using bio-signals. The design of the experiments addresses the research questions by systematically analyzing different aspects of model performance, the effectiveness of ensemble learning, and the influence of multi-modal bio-signal integration.

Dataset Splitting Three subsets of the data were created to guarantee an all-around assessment of model performance:

- **Training Set:**applied for weight optimization and model training. The learning process is mostly formed by this collection, which helps the models to discover patterns and relationships inside the data.
- **Validation Set:** Used to stop overfitting and for hyperparameter adjustment. Fine-tuning model parameters made possible by the validation set lets one improve performance on fresh data.
- **Test Set:** Set aside for last review of model performance. The test set offers an objective evaluation of the models' capacity for generalizing to fresh, untested data.

Stratified division of the data guarantees that every class is proportionately represented in every subset. This guarantees that the models are trained and evaluated on a varied spectrum of samples and helps to preserve the balance of the dataset.

Model Types Several machine learning models were created and assessed to forecast smoking behavior. The popularity and efficiency of these models in binary classification problems helped to guide their selection. The experimentally utilized models include:

- **Logistic Regression:**
  - **Solver:** lbfgs
  - **Penalty:** L2 regularization to avoid overfitting
  - **Max Iterations:** 1000 to ensure convergence
  - **Note:** Applied in both individual and stacked ensemble environments.
- **Random Forest:**
  - **Criterion:** gini
  - **Number of Estimators:** 100
  - **Max Depth:** None.
- **LightGBM:**
  - **Boosting Type:** gbdt (Gradient Boosting Decision Tree)
  - **Learning Rate:** 0.1
  - **Number of Leaves:** 31.
- **CatBoost:**
  - **Learning Rate:** 0.03
  - **Loss Function:** Logloss.

- **XGBoost:**
  - **Booster:** gbtree, using gradient boosting trees
  - **Learning Rate:** 0.1
  - **Max Depth:** 6.
- **Stacking Meta-Classifier:**
  - **Base Models:** Logistic Regression, Random Forest, LightGBM, CatBoost, XGBoost
  - **Final Estimator:** Logistic Regression.
  - **Cross-Validation:** The stacking model's stability was evaluated using 10-fold cross-validation.

Every model was trained with its own hyperparameters, and its performance was assessed with a consistent set of criteria thereby enabling a fair comparison.

Training and Evaluation The training procedure was carefully designed to prevent overfitting and guarantee the models acquired efficient knowledge from the data:

- **Individual Models:** Every model was hyperparameter tuned on the training set by means of grid search. The models were validated and overfitting was avoided by means of cross-validation. The last model was subsequently tested on a set.
- **Ensemble Method (Stacking Classifier):** Input features for the stacking classifier were the predictions from each individual model. This combined strategy sought to offset the individual shortcomings of every model by using their strengths. Ten-fold cross-valuation was used to train the stacking classifier so as to guarantee robustness and stop overfitting.

Control Variables Several variables were under control all through the research to guarantee that the given studies yielded consistent and trustworthy results:

- **Hyperparameters:** Grid search and hand tuning helped each model type fine-tune key hyperparameters including learning rate, number of estimators, and max depth.
- **Model Architecture:** For ensemble techniques such as the stacking classifier, the architecture was meticulously crafted to guarantee that the underlying models significantly influenced the resultant prediction.
- **Class Imbalance:** Class imbalance was investigated in the dataset, and methods including stratified sampling and class weighting were used to guarantee the models did not become biassed toward the majority class.
- **Evaluation Metrics:** Accuracy, F1 score, and AUC-ROC defined the models' evaluation. These measures were selected to offer a whole picture of model performance, thereby including class discrimination ability as well as precision and recall.

Summary This experimental setup was meticulously designed to evaluate the impact of different model types on the prediction of smoking behavior using bio-signals. The setup ensures that the models are rigorously tested across various conditions, providing valuable insights into their strengths and weaknesses. The consistent application of evaluation metrics and control variables ensures that the results are reliable and comparable across different models.

Table 6.3: Hyperparameter settings used for the ensemble models in this study.

Hyperparameter Settings	
Learning Rate	0.001
Batch Size	32
Epochs	100
Optimizer	Adam
Early Stopping	10 epochs
Random Forest Estimators	100
Random Forest Max Depth	8
XGBoost Max Depth	6
LightGBM Max Depth	6
CatBoost Iterations	500
Logistic Regression Solver	'liblinear'
Samples in Training Set	127404
Samples in Validation Set	31851

6.2.2 Experiment Results

This part explores the two main research issues raised in this project, therefore offering a thorough analysis of the results of many experiments. These carefully planned tests investigate the benefits of adopting an ensemble stacking technique and evaluate the predictive performance of individual machine learning models in the framework of smoking behavior detection using bio-signals.

6.2.3 Research Question 1 (RQ1): How do individual machine learning models perform in predicting smoking behaviors using bio-signals?

Answer to RQ1:

A set of well-known algorithms was chosen in order to fully explore the effectiveness of particular machine learning models in smoking behavior prediction. Test models were Logistic Regression, Random Forest, XGBoost, LightGBM, and Catboost. Every one of these models has unique qualities that fit various facets of categorization assignments. Using important criteria including accuracy, F1 score, and AUC-ROC, which taken together offer a whole picture of the models' predictive capacity, the performance of these models was thoroughly assessed.

1. Logistic Regression

Because logistic regression is so simple and used so extensively in binary classification problems, it was selected as the baseline model. This model is a linear classifier, so it assumes in this case smoking status a linear relationship between the input features and the log-odds of the target variable. Logistic Regression, for all its simplicity, can be rather successful in cases with almost linear relationships between the variables. The Logistic Regression model in this work obtained an accuracy of 75.02%, an F1 score of 72.67%, and an AUC-ROC of 0.76. With 75.02% of the test set correctly identified, the accuracy shows that the model Balancing accuracy and recall, the F1 score indicates that the model has a respectable capacity to find smokers while reducing false positives and false negatives. With a 0.76 AUC-ROC score—a gauge of the model's ability to discriminate between classes—the discriminative power was rather modest.

Conclusion: Easy to understand and a dependable baseline model, logistic regression offered. Its capacity to detect the intricate patterns in the bio-signal data is limited, nonetheless, by its dependence on linear connections between the characteristics and the target variable. Consequently, even if it operates satisfactorially, it lacks the intricacy needed to reach the best degrees of accuracy and dependability in this use. Strength of the model are its simplicity and interpretability; but, these come at the expense of lower performance in more complicated situations.

2. Random Forest

Expected to better manage the non-linearities and interactions between features than Logistic Regression, Random Forest is an ensemble learning method that creates several decision trees during training and outputs the mode of the classes (classification) of the individual trees. With an accuracy of 74.83%, an F1 score of 71.55%, and an AUC-ROC of 0.75 the Random Forest model attained While Random Forest is better at identifying non-linear associations, the slight drop in accuracy relative to Logistic Regression points to possible overfitting, especially in high-dimensional data environments like this. When a model learns the noise in the training data instead of the real signal, overfitting results, which reduces performance on new data.

Conclusion: Random Forest showed a reasonable capacity to replicate non-linear interactions between the smoking status and the bio-signal characteristics. Its usefulness was hampered, nonetheless, by its inclination to overfit—especially in the presence of many features. Although the model's ensemble character usually offers robustness, in this case it struggled to generalize as well as planned most likely because of the complexity and variety of the bio-signal data.

3. XGBoost

Extreme Gradient Boosting, or XGBoost, is well-known for both performance on a range of classification challenges and efficiency. One tree builds after another, each one seeking to fix the mistakes of the one before it. Popular for difficult datasets, XGBoost also features a regularizing method to avoid overfitting. In this work, XGBoost achieved 79.34% accuracy, 76.12% F1 score, and 0.86 AUC-ROC. These results show XGBoost more deftly handling the complexity of the bio-signal data than Random Forest and Logistic Regression. Since XGBoost displays its greater ability to distinguish between smokers and non-smokes, the high AUC-ROC score indicates that XGBoost effectively detected the fundamental trends in the data.

Conclusion: XGBoost has proven resilient and capable of handling difficult datasets that contributed to its superior performance in this exercise. It was successful because it had the ability to penalize and regularize your model, handle missing data, simulate intricate interaction among variables and save you from overfitting! As the results show, XGBoost performs really well to such complex predictive modeling problems with non-linear relationships between features.

4. LightGBM

LightGBM, commonly known as Light Gradient Boosting Machine, is another gradient boosting tool that generates trees depending on histogram-based methods, therefore enabling faster training and more efficient memory utilization. Since LightGBM is particularly fit for these settings, it is a strong choice for this work particularly in large datasets with many features. With an accuracy of 79.21%, an F1 score of 75.89%, and an AUC-ROC of 0.86 the model performed. These findings show that LightGBM is equally successful in this environment since they are quite similar to those achieved with XGBoost. The somewhat lower F1 score implies that, although LightGBM trained quickly, it might have sacrificed some accuracy or recall relative to XGBoost.

Conclusion: From bio-signals, LightGBM showed to be a strong and quick model for estimating smoking behavior. Especially in situations when training speed is a top concern, its capacity to manage big datasets rapidly without a notable performance loss makes it appealing substitute for XGBoost. The performance of LightGBM in this work validates its ranking as a top-notional model for difficult classification problems.

5. CatBoost

Designed especially to naturally handle categorical features natively, CatBoost—also known as Categorical Boosting—does not need extensive preprocessing like one-hot encoding. Working with datasets including multiple category variables will especially help from this function. CatBoost is also appropriate for datasets including a mix of feature kinds since it uses methods to fight overfitting. With an accuracy of 79.68%, an F1 score of 76.45%, and an AUC-ROC of 0.87. CatBoost came out as the top individual model in this work. These measures show that CatBoost not only efficiently managed the bio-signal data but also performed very well in spotting the subtleties between smokers and non-smoking individuals. The high AUC-ROC score indicates outstanding discriminative ability; the F1 score shows a good balance between recall and precision.

Conclusion: CatBoost had a small advantage over the other models examined because of its unusual treatment of categorical data and strong avoidance of overfitting. It is the best performer in this research since it excels in all important criteria. The results of CatBoost show its capacity to provide extremely accurate predictions and properly represent complicated, mixed-type datasets.

Overall Conclusion for RQ1

Using bio-signals, the performance of the individual models varied; CatBoost and XGBoost emerged as the most successful algorithms for smoking behavior prediction. These models exceeded others because of their increased capacity to manage categorical variables, control overfitting, and handle complicated relationships between characteristics. Although Random Forest and Logistic Regression offered good baselines, their shortcomings in managing the complexity of the data were clear. Though Catboost's tailored approach to category data processing finally gave it the advantage, LightGBM also performed brilliantly and closely matched XGBoost. These results emphasize the need of choosing models not only strong but also customized to the particular features of the data.

6.3 Research Question 2 (RQ2): Does a stacking classifier combining multiple ensemble models (Logistic Regression, Random Forest, LightGBM, CatBoost, XGBoost) improve prediction accuracy for smoking habits compared to individual models?

6.3.1 Answer to RQ2

Combining many distinct machine learning models allowed an ensemble stacking classifier to be created to fully handle RQ2. Using a stacking ensemble makes sense since it allows one to maximize the strengths of many models, so reducing their individual flaws.

Stackings of models allow the ensemble to theoretically attain accuracy, resilience, and generalization better than any one model could achieve on its own. Each of the models used for this stacking method—Logistic Regression, Random Forest, XGBoost, LightGBM, and Catboost—contributes special value to the ensemble.

Model Selection and Ensemble Strategy Selecting five different machine learning models, each with a different method of handling the data, the stacking ensemble model was carefully built. Included for its simplicity and interpretability, logistic regression was a linear model able to rapidly identify and balance the most pertinent features. Random Forest was selected for its capacity to capture non-linear correlations using its ensemble of decision trees, which taken collectively provide a strong means of preventing overfitting on the training data. XGBoost presented its strong gradient boosting method, which uses regularization to enhance model generalization and consecutively fixes mistakes of past models. Particularly with big datasets, LightGBM was added for its efficiency and speed; it is a great match to XGBoost with a similar, but different, boosting technique. CatBoost was chosen for its advanced overfitting prevention methods, which are absolutely essential in preserving the generalizability of the model, and for its exceptional handling of categorical data.

Stacking Mechanism These foundation models—each trained on the same dataset—form the first layer in a stacking ensemble. These models independently project on the training data; their outputs are subsequently fed into a meta-learner. In this situation, the meta-learner was logistic regression. The meta-learner's job is to learn how to optimally combine the predictions from the base models, therefore producing a final model with combined strengths of all the base learners. The training approach split the data into several folds, whereby each base model was trained on one of the several folds and their predictions were aggregated. This method guarantees that the stacking model improves generalizing capacity over several data scenarios and is not overfitting to any one dataset split.

Performance Evaluation

Three main criteria—accuracy, F1 score, and AUC-ROC—were thoroughly assessed for the stacking ensemble. These measures give a comprehensive evaluation of the accuracy with which the model classifies smoking habits. Accuracy of the stacking classifier was 78.03%. About 78% of the test dataset's occurrences had the ensemble accurately forecast their smoking status. Using several learning techniques clearly shows the benefit since this performance is noticeably better than the accuracy of most individual models. The stack's F1 score came out to be 75.98%. The F1 score strikes a mix between accuracy and recall, revealing how well the model can appropriately spot smokers (positive class) while lowering false positives and false negatives. The F1 score of the stacking model exceeded that of the separate models, so underscoring its enhanced capacity in managing the subtleties of smoking behavior classification. The stacker's AUC-ROC came out at 0.87. This statistic assesses, over all classification thresholds, the model's capacity to discriminate between the smoker and non-smoking classes. Strong across several choice thresholds, a high AUC-ROC score indicates that the stacking ensemble efficiently balances the trade-off between sensitivity (recall) and specificity (true negative rate).

Comparative Analysis with Individual Models

Clearly showing better performance than the individual models was the stacking ensemble. Though strong in their own right, each model had particular restrictions. Although logistic regression was simple and lacked the capacity to catch intricate non-linear patterns, which so reduced its general accuracy and robustness. Although Random Forest was more adept at capturing non-linear interactions, it was prone to overfitting, therefore compromising its generalizing power. Though their performance was somewhat limited by the constraints inherent in each method, XGBoost and LightGBM offered great accuracy and were adept at managing complicated interactions inside the data. CatBoost had the best single performance overall because it performed well in handling categorical variables and avoiding overfitting, but not to beat ensemble. The stacking classifier was able to deal with these constraints by logically combing strengths from both models. For Logistic Regression, the linearity of prediction is compensated by non-linear modeling capabilities reminiscent of Random Forest and the boosting strength exhibited in models like XGBoost, LightGBM and Catboost. This integration led to a better trade-off model for keeping up with different data distributions.

Robustness and Generalization

Among the key advantages of the stacking ensemble was robustness. The generalizing capacity of the model to unprocessed data was much enhanced by combining the projections from different techniques. This robustness is reflected in the strong AUC-ROC score, which indicates that the model performed well over numerous thresholds and was less prone to be influenced by data noise or outliers. Moreover, the performance of the ensemble on numerous criteria indicates its decreased dependence on any one quality or pattern in the data, thereby raising its reliability in practical uses.

Overall Conclusion for RQ2

The experiments using bio-signals amply demonstrated, in terms of smoking behavior prediction, the performance of the stacking classifier above all others. Emphasizing the advantages of combining numerous models for challenging prediction tasks, the ensemble method presented a better balance of accuracy, F1 score, and AUC-ROC. Especially, the stacking technique generated a more strong and generally applicable prediction model by combining the features of every base model. This outcome validates the hypothesis that a stacking ensemble will outperform single models, thereby guiding strategy for similar predictive modeling issues in future.

6.4 Model Generalization

Model generalization in machine learning is the ability of a model to perform adequately on new, unseen data not contained in the training set. This capacity is essential to guarantee that the model learns fundamental patterns applicable to a larger spectrum of data points instead of merely memorizing the training data. In this work, model generalization was thoroughly assessed to guarantee that the forecasts about smoking behavior depending on bio-signals would be dependable and relevant in practical situations.

6.4.1 Evaluation of Generalization

Many approaches were used to evaluate the generalization of the proposed models in this work. K-fold cross-validation, in which the dataset was split into several subsets and the model was trained and assessed methodically over several subsets, was the main method applied to assess generalization. This approach indicated that the models did not rely too much on certain patterns in the training data as they maintained constant performance across many subsets of data. Particularly showing strong generalization with little variance in performance across the several folds was the stacking classifier.

Ensemble Learning Generalization was much enhanced by ensemble learning, particularly with reference to the stacking classifier. Ensemble techniques are well-known for their ability to average the outputs of numerous models, therefore lowering the variation of forecasts. In this work, a model that generalized well over several subsets of the data was produced by aggregating the strengths of several base models, each trained with regularization and cross-validation. This method not only raised accuracy but also produced more consistent and trustworthy forecasts, therefore lowering the possibility of overfitting to any one component of the training data.

Validation Set Performance The models were tested on a separate validation set not utilized during training to replicate how they would perform on totally fresh data following training. Further proving the models' capacity to generalize, the validation set findings tightly matched the cross-validation results. Particularly, the stacking classifier kept good accuracy and F1 scores on the validation set, suggesting that it was fairly faithfully capturing the underlying trends connected to smoking habits instead of overfitting to the training data.

Generalization Across Different Bio-Signals The study also looked at how the model performed while aggregating several bio-signals. Integrating several signals allowed the model to access a wider spectrum of physiological data, hence improving its capacity to generalize over many kinds of inputs. Strong generalization and excellent accuracy sustained across several subsets of the data were displayed by the combined bio-signal model. This result implies that multi-modal methods may greatly improve the generalization and resilience of machine learning models, hence increasing their relevance in practical environments.

6.4.2 Challenges and Considerations

Although the models showed great generalizing capacity, numerous difficulties arose. The dataset utilized in the study might have had an imbalance in the distribution of smokers and non-smoking individuals, which can influence the generalizing capacity of the model since it might get biased toward the majority class. Resampling and class weighting were among the methods thought to help to reduce this problem and enhance generalization. Generalization depends partly on the ability to capture intricate interactions among characteristics. Although the stacking classifier improved in this sense, the model might still miss minor interactions. To improve generality even more, future research might look at more sophisticated methods such as deep learning.

6.4.3 Conclusion

Based on bio-signals in real-world situations, the models created in this study showed great generalization capacity, which qualifies them for estimating smoking behavior. Regularization methods, ensemble learning, cross-validation, and multi-modal data integration used together showed to produce models that not only performed well on the training data but also on fresh, unseen data. This strong generalization guarantees that the models can be boldly used in practical applications, such as health monitoring and smoking cessation programs, where dependability of predictions is essential for success.

7 Conclusion

This work reveals that integrating bio-signals with advanced machine learning techniques can greatly increase the accuracy and resilience of forecast models for smoking behavior. Using a stacking classifier—which aggregates the strengths of many ensemble models, including Logistic Regression, Random Forest, LightGBM, Catboost, and XGBoost—then obtained higher prediction accuracy than using individual models. This work underlines the importance of multi-signal integration by showing that combining numerous bio-signals, produces a more strong and accurate framework for behavioral prediction. Moreover, the application of Local Interpretable Model-agnostic Explanations (LIME) improved understanding of model predictions, so facilitating open analysis on the factors influencing smoking behavior. This work not only enhances the current methods of smoking habit identification but also generates possibilities for applying similar techniques in other domains of healthcare and behavioral analytics. The work highlights generally the possibility of effectively solving difficult behavioral prediction issues by merging powerful machine learning models with bio-signals.

7.1 Summary

This work has focused over the past three months on creating a thorough and strong model for bio-signals-based smoking behavior prediction. The effort began with selecting a good dataset from Kaggle with pertinent bio-signals. The basis for later feature engineering, model building, and evaluation cycles was this dataset.

7.1.1 Analysis of Available Technologies

The first step consisted on a careful study of current bio-signal processing and machine learning technologies, namely in relation to behavioral prediction. This study comprised a review of modern approaches applied in smoking behavior prediction as well as the present literature. The aim was to find the most efficient methods and instruments that might be used to raise the accuracy and robustness of smoking habit forecasts.

7.1.2 Design and Implementation on Windows OS

Designs and implementations took place on a Windows platform. This included selecting feature engineering techniques and building a machine learning pipeline featuring a stacking classifier. Combining many ensemble models—Logistic Regression, Random Forest, LightGBM, Catboost, and XGBoost—the stacking classifier tried to maximize their strengths and increase general prediction accuracy. This phase also included extensive testing and validation to ensure the resilience and dependability of the model. These experiments greatly helped the second study issue, which sought whether a stacking classifier provided better prediction accuracy than single models.

7.1.3 Documentation of the Process

Every stage of the process was recorded using complete documentation retained all through the project. Apart from the results of various tests and assessments, this material included technical justifications of the employed feature engineering methods and machine learning models. The material proved to be a helpful tool for direction of upcoming research in this sector and for guaranteeing the reproducibility of the work.

7.1.4 Evaluation of the Proposed Solution

A comprehensive evaluation of the proposed solution under several parameters including ROC AUC, F1 score, and accuracy constituted the last phase. The model reportedly significantly improved robustness and prediction accuracy above existing techniques. Particularly, the stacking classifier exceeded single models to get the best overall metrics. This analysis addressed the third study question, which investigated the effects of aggregating several bio-signals on model accuracy and resilience, and offered understanding of the efficacy of the multi-modal bio-signal strategy.

The main turning points reached on this project are summed up here. Using bio-signals, the researchers produced a quite accurate and interpretable model for estimating smoking behavior. The knowledge acquired from this effort not only develops the discipline of healthcare analytics but also offers a basis for next studies and uses in behavioral prediction and individualized health monitoring.

7.2 Dissemination

Academic research as well as commercial applications in the sectors of behavioral analytics and healthcare should benefit much from the model and methods used in this work.

7.2.1 Potential Users

Professionals working in the domains of healthcare analytics, behavioral science, and bioinformatics as well as researchers would be the main consumers of this component. In public health campaigns, where early identification of smoking behavior can guide preventive and intervention plans, the model's capacity to forecast smoking behavior based on bio-signals could be especially helpful. By spotting patients at risk of smoking-related ailments, the model might also be included into more general health monitoring systems used by hospitals and clinics, therefore contributing to individualized healthcare.

7.2.2 Industry and Research Projects

Regarding industry applicability, businesses focused in wearable health technology and digital health platforms could find use for the model. To offer real-time health information, these sectors are progressively emphasizing including machine learning models into their solutions. The model created here could be included into wearable devices tracking bio-signals that provide users feedback on their health practices and maybe inspire better lifestyles by means of their interaction.

Moreover, the approach might be included into EU initiatives aiming at public health and behavior modification. Large-scale initiatives aiming at lowering smoking prevalence across Europe could, for example, use this model to better grasp the elements impacting smoking behavior in various groups. Such combination would improve the capacity of the project to focus interventions more precisely depending on real-time data.

7.2.3 Open Source and Broader Integration

Furthermore supplied as an open-source tool would be the model, which would let the larger research community expand upon it. Other researchers could improve and alter the model for various use cases, such predicting other health behaviors or merging extra bio-signals, by sharing the model and its underlying code as an open-source project. Along with advancing scientific understanding, this would encourage multidisciplinary cooperation.

From improving personal health monitoring to supporting major public health campaigns, this study is likely to have a significant influence in various spheres. By means of distribution in both academia and industry, this model has great potential to enhance health outcomes and forward scientific knowledge of behavioral prediction.

7.3 Problems Encountered

Throughout the course of this research, several challenges were encountered, primarily related to the limitations of computational resources and time constraints.

7.3.1 Computational Limitations

One of the significant challenges faced was the attempt to implement SHAP (SHapley Additive exPlanations) and Partial Dependence Plots (PDP) for model interpretation. These methods are computationally intensive, especially when applied to large datasets and complex ensemble models. Due to the limited computing power of the system used in this research, running SHAP and PDP for all the models proved to be unfeasible. The computational load not only caused significant delays but also led to incomplete or failed executions. Consequently, these advanced interpretability techniques could not be fully integrated into the final analysis. While LIME provided valuable insights, the absence of SHAP and PDP means that some aspects of global model interpretation and feature interaction analysis could not be explored in depth.

7.3.2 Integration into Real-Time Applications

Another challenge was the intended integration of the model into a real-time application for continuous monitoring and prediction of smoking behavior based on bio-signals. This integration would have required not only a robust, real-time data processing pipeline but also a high level of optimization to ensure the model could run efficiently on various devices, such as smartphones or wearable technology. However, due to the three-month timeframe allocated for this research, it was not feasible to develop and test such an application. The complexities involved in real-time data handling and the optimization of machine learning models for mobile and embedded systems were beyond the scope of this study given the limited time.

7.3.3 Time Constraints

The three-month duration of this project also posed a significant challenge. While substantial progress was made in developing and evaluating the predictive model, the limited time meant that certain aspects of the research, such as deeper exploration of multi-signal integration and more sophisticated model tuning, could not be fully pursued. This time constraint also limited the ability to conduct broader testing across different datasets to improve the generalizability of the findings.

Notwithstanding these difficulties, the work effectively showed how well advanced feature engineering and ensemble learning models might predict smoking behavior. The difficulties faced, however, point to areas for future development and imply that these challenges could be surmounted given additional time and money.

7.4 Outlook

Looking ahead, there are various interesting paths for expanding this study, especially in resolving the constraints faced and investigating fresh directions for improving the applicability and robustness of the model.

7.4.1 Enhancing Interpretability with SHAP and PDP

Successful integration of SHAP and PDP into the model interpretability framework is one of the main directions of future development. Understanding complicated models like the stacking ensemble utilized in this work depends on thorough knowledge of feature importance and interactions, which these techniques offer. These methods can be completely applied with more strong computer resources, therefore enabling a closer investigation of how particular traits influence model predictions. This would improve the model's openness as well as offer medical professionals additional practical information.

7.4.2 Real-Time Application and Deployment

Development of a real-time application for tracking smoking habit using bio-signals represents yet another crucial focus for further research. This would entail merging the model with real-time data streams and maximizing it for use on mobile devices. This is a realistic target given developments in model compression methods and rising processing capability of mobile devices. In public health and individualized healthcare, this kind of application could offer major advantages since it could give early warning systems for smoking behavior and constant monitoring.

7.4.3 Broader Testing and Generalization

Future studies should include testing the model on several datasets, including those from various demographic and geographic areas, thereby improving the generalizability of the model. This guarantees the model's applicability in a larger spectrum of environments and helps one to grasp its performance in many circumstances. Further enhancing the predicted accuracy and resilience of the model would be extending it to include other bio-signals and outside data sources, such genetic information or environmental elements.

7.4.4 Exploration of New Interpretability Techniques

Nikam et al. demonstrated via their explainable technique for species identification using LIME, which stresses the method's relevance in generating intelligible model predictions, boosting the transparency of machine learning models—especially in challenging identification tasks—is vitally crucial. [NRP+22]. Beyond SHAP and PDP, exploring other interpretability techniques, such as counterfactual explanations, could provide new perspectives on model behavior. For instance, counterfactual explanations might assist one grasp what little variations in the input attributes would produce different prediction results. This could be particularly useful in scenarios where it is important to understand the thresholds or tipping points that influence smoking behavior predictions.

7.4.5 Expanding the Scope of Applications

Finally, while this research focused on smoking behavior prediction, the methodologies developed could be applied to other areas of behavioral and health analytics. Future work could explore the applicability of this approach to other health behaviors, such as diet, physical activity, or substance use. By adapting the model to different contexts, it could contribute to a broader understanding of how bio-signals can be used to predict and influence a wide range of health outcomes.

7.5 Problems Encountered

7.5.1 Computational Limitations

Implementing SHAP (SHapley Additive exPlanations) and Partial Dependence Plots (PDP) for model interpretation proved one of the main difficulties. Particularly in relation to large datasets and sophisticated ensemble models, these techniques are computationally demanding. Running SHAP and PDP for all the models proved to be impossible given the system's restricted computational capability used in this work. Apart from major delays, the computational burden resulted in partial or failed executions. As so, these sophisticated interpretability methods could not be completely included into the last study. Although LIME gave insightful analysis, some features of global model interpretation and feature interaction analysis could not be thoroughly investigated in lack of SHAP and PDP.

7.5.2 Integration into Real-Time Applications

The anticipated integration of the model into a real-time application for continuous monitoring and prediction of smoking behavior depending on bio-signals presented still another difficulty. This integration would have needed not just a strong, real-time data processing pipeline but also high degrees of optimization to guarantee the model could function effectively on many devices, such wearable technologies or cellphones. Nevertheless, it was not possible to create and test such an application given the three-month duration allotted for this study. Given the restricted time, the complexity of real-time data handling and the optimization of machine learning models for mobile and embedded systems beyond the purview of this study.

7.5.3 Time Constraints

Another great difficulty of this three-month project was Although the predictive model was developed and evaluated with great progress, the restricted time meant that several areas of the research, such deeper investigation of multi-signal integration and more advanced model tuning, could not be totally pursued. This time limit also restricted the capacity to do more extensive testing over several datasets, hence enhancing the generalizability of the results.

Notwithstanding these obstacles, the study effectively showed how well advanced feature engineering and ensemble learning models might predict smoking behavior. The difficulties faced, however, point to areas for future development and imply that these challenges could be surmounted given additional time and money.

7.6 Outlook

Looking ahead, especially in addressing the constraints faced and investigating fresh paths for improving the applicability and robustness of the model, there are various interesting prospects for extending this study.

7.6.1 Enhancing Interpretability with SHAP and PDP

Successful integration of SHAP and PDP into the model interpretability framework is one of the main directions of future development. Understanding complicated models like the stacking ensemble utilized in this work depends on thorough knowledge of feature importance and interactions, which these techniques offer. These methods can be completely applied with more strong computer resources, therefore enabling a closer investigation of how particular traits influence model predictions. This would improve the model's openness as well as offer healthcare professionals additional practical information.

7.6.2 Real-Time Application and Deployment

Development of a real-time application for tracking smoking habit using bio-signals represents yet another crucial focus for further research. This would entail merging the model with real-time data streams and maximizing it for use on mobile devices. This is a realistic target given developments in model compression methods and rising processing capability of mobile devices. In public health and individualized medicine, such an application could offer major advantages by providing early warning systems for smoking behavior and continual monitoring.

7.6.3 Broader Testing and Generalization

Future studies should include testing the model on several datasets, including those from various demographic and geographic areas, thereby improving the generalizability of the model. This guarantees the model's applicability in a larger spectrum of environments and helps one to grasp its performance in many circumstances. Further enhancing the predicted accuracy and resilience of the model would be extending it to include other bio-signals and outside data sources, such genetic information or environmental elements.

7.6.4 Expanding the Scope of Applications

At last, even if this study concentrated on smoking behavior prediction, the created approaches could find utility in various spheres of behavioral and health analytics. Future research could investigate the relevance of this strategy to other health practices such physical exercise, diet, or substance use. Adapting the model to various settings could help to clarify how bio-signals might be utilized to forecast and affect a wide spectrum of health outcomes.

In essence, there are many chances for future growth even if this study has achieved great progress in integrating bio-signals and advanced machine learning approaches for smoking behavior prediction. Future research can build on these roots to provide more strong, interpretable, and generally applicable models by tackling the difficulties faced and investigating new directions.

List of Acronyms

AI Artificial Intelligence

BMI Body Mass Index

CatBoost Categorical Boosting

CDC Centers for Disease Control and Prevention

COPD Chronic Obstructive Pulmonary Disease

ECG Electrocardiograms

EDA Exploratory Data Analysis

EEG Electroencephalograms

EMG Electromyograms

HDL High-Density Lipoprotein

HRV Heart Rate Variability

KNN K-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

LDL Low-Density Lipoprotein

LIME Local Interpretable Model-agnostic Explanations

LightGBM Light Gradient Boosting Machine

ML Machine Learning

PDP Partial Dependence Plots

ROC AUC Area Under the Receiver Operating Characteristic Curve

SHAP SHapley Additive exPlanations

WHO World Health Organization

XGBoost eXtreme Gradient Boosting

Bibliography

[AESI+20] Ali, Farman, Shaker El-Sappagh, SM Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran and Kyung-Sup Kwak: *A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion.* Information Fusion, 63:208–222, 2020.

[AFG+21] Abukmeil, Mohanad, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri and Fabio Scotti: *A survey of unsupervised generative models for exploratory data analysis and representation learning.* Acm computing surveys (csur), 54(5):1–40, 2021.

[AJ24] Al-Jamimi, Hamdi A: *Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction.* IEEE Access, 2024.

[ANM22] Abdollahi, Jafar and Babak Nouri-Moghaddam: *Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction.* Iran Journal of Computer Science, 5(3):205–220, 2022.

[AT21] Abo-Tabik, Maryam A: *Using Deep Learning Predictions of Smokers' Behaviour to Develop a Smart Smoking-Cessation App.* PhD thesis, Manchester Metropolitan University, 2021.

[ATCDB19] Abo-Tabik, Maryam, Nicholas Costen, John Darby and Yael Benn: *Decision tree model of smoking behaviour.* In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1746–1753. IEEE, 2019.

[BLY+18] Burnham, Jason P, Chenyang Lu, Lauren H Yaeger, Thomas C Bailey and Marin H Kollef: *Using wearable technology to predict health outcomes: a literature review.* Journal of the American Medical Informatics Association, 25(9):1221–1227, 2018.

[Cen23] Centers for Disease Control and Prevention: *About Tobacco*, 2023. Accessed: 2023-08-26.

[CFN+21] Chriskos, Panteleimon, Christos A Frantzidis, Christiane M Nday, Polyxeni T Gkivogkli, Panagiotis D Bamidis and Chrysoula Kourtidou-Papadeli: *A review on current trends in automatic sleep staging through bio-signal recordings and future challenges.* Sleep medicine reviews, 55:101377, 2021.

[Cho20] Cho, Jae Hyuk: *Detection of smoking in indoor environment using machine learning.* Applied Sciences, 10(24):8912, 2020.

[CJDR20] Caccamisi, Andrea, Leif Jørgensen, Hercules Dalianis and Mats Rosenlund: *Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records.* Upsala journal of medical sciences, 125(4):316–324, 2020.

[CJF$^+$21] Choi, Jeeyae, Hee-Tae Jung, Anastasiya Ferrell, Seoyoon Woo and Linda Haddad: *Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (hookah) users.* Journal of Clinical Medicine, 10(5):972, 2021.

[CJS98] Corbett, PWM, JL Jensen and KS Sorbie: *A review of up-scaling and cross-scaling issues in core and log data interpretation and prediction.* Geological Society, London, Special Publications, 136(1):9–16, 1998.

[CRDLTD$^+$22] Chaganti, Rajasekhar, Furqan Rustam, Isabel De La Torre Díez, Juan Luis Vidal Mazón, Carmen Lili Rodríguez and Imran Ashraf: *Thyroid disease prediction using selective features and machine learning techniques.* Cancers, 14(16):3914, 2022.

[CSCMP$^+$20] Caiafa, Cesar Federico, Jordi Solé-Casals, Pere Marti-Puig, Sun Zhe and Toshihisa Tanaka: *Decomposition methods for machine learning with small, incomplete or noisy datasets.* Applied Sciences, 10(23):8481, 2020.

[CWC$^+$22] Chiu, Chih-Chou, Chung-Min Wu, Te-Nien Chien, Ling-Jing Kao, Chengcheng Li and Han-Ling Jiang: *Applying an improved stacking ensemble model to predict the mortality of ICU patients with heart failure.* Journal of Clinical Medicine, 11(21):6460, 2022.

[DLPR20] Davagdorj, Khishigsuren, Jong Seol Lee, Van Huy Pham and Keun Ho Ryu: *A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention.* Applied Sciences, 10(9):3307, 2020.

[DPTUR20] Davagdorj, Khishigsuren, Van Huy Pham, Nipon Theera-Umpon and Keun Ho Ryu: *XGBoost-based framework for smoking-induced noncommunicable disease prediction.* International journal of environmental research and public health, 17(18):6513, 2020.

[DYC$^+$20] Dong, Xibin, Zhiwen Yu, Wenming Cao, Yifan Shi and Qianli Ma: *A survey on ensemble learning.* Frontiers of Computer Science, 14:241–258, 2020.

[FKM$^+$23] Fu, Rui, Anasua Kundu, Nicholas Mitsakakis, Tara Elton-Marshall, Wei Wang, Sean Hill, Susan J Bondy, Hayley Hamilton, Peter Selby, Robert Schwartz et al.: *Machine learning applications in tobacco research: a scoping review.* Tobacco Control, 32(1):99–109, 2023.

[FKT$^+$23] Fatima, Ruhi, Sabeena Kazi, Asifa Tassaddiq, Nilofer Farhat, Humera Naaz and Sumera Jabeen: *Stacking Ensemble Machine Learning Algorithm with an Application to Heart Disease Prediction.* Contemporary Mathematics, pages 905–925, 2023.

[GAA+22] Gollapalli, Mohammed, Aisha Alansari, Heba Alkhorasani, Meelaf Alsubaii, Rasha Sakloua, Reem Alzahrani, Mohammed Al-Hariri, Maiadah Alfares, Dania AlKhafaji, Reem Al Argan et al.: *A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM.* Computers in Biology and Medicine, 147:105757, 2022.

[GLB+23] Ghasemieh, Alireza, Alston Lloyed, Parsa Bahrami, Pooyan Vajar and Rasha Kashef: *A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients.* Decision Analytics Journal, 7:100242, 2023.

[HELMGA20] Hao, Tianxiao, Jane Elith, José J Lahoz-Monfort and Gurutzeta Guillera-Arroita: *Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models.* Ecography, 43(4):549–558, 2020.

[HK20] Hancock, John T and Taghi M Khoshgoftaar: *Survey on categorical data for neural networks.* Journal of big data, 7(1):28, 2020.

[HM21] Hellen, Nakayiza and Ggaliwango Marvin: *Interpretable feature learning framework for smoking behavior detection.* arXiv preprint arXiv:2112.08178, 2021.

[HMR+23] Huang, FeiMing, QingLan Ma, JingXin Ren, JiaRui Li, Fen Wang, Tao Huang and Yu-Dong Cai: *Identification of Smoking-Associated Transcriptome Aberration in Blood with Machine Learning Methods.* BioMed research international, 2023(1):5333361, 2023.

[HQS+20] Hu, Zhixu, Hang Qiu, Ziqi Su, Minghui Shen and Ziyu Chen: *A stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases.* IEEE Access, 8:138719–138729, 2020.

[HYC+22] Huang, Weiting, Tan Wei Ying, Woon Loong Calvin Chin, Lohendran Baskaran, Ong Eng Hock Marcus, Khung Keong Yeo and Ng See Kiong: *Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction.* Scientific Reports, 12(1):1033, 2022.

[JPK+14] Jang, Eun-Hye, Byoung-Jun Park, Sang-Hyeob Kim, Youngji Eum and Jin-Hun Sohn: *A study on analysis of bio-signals for basic emotions classification: recognition using machine learning algorithms.* In *2014 International Conference on Information Science & Applications (ICISA)*, pages 1–4. IEEE, 2014.

[KGG21] Kalagotla, Satish Kumar, Suryakanth V Gangashetty and Kanuri Giridhar: *A novel stacking technique for prediction of diabetes.* Computers in Biology and Medicine, 135:104554, 2021.

[LDB+22] Laatifi, Mariam, Samira Douzi, Abdelaziz Bouklouz, Hind Ezzine, Jaafar Jaafari, Younes Zaid, Bouabid El Ouahidi and Mariam Naciri: *Machine learning approaches in Covid-19 severity risk prediction in Morocco.* Journal of big Data, 9(1):5, 2022.

[LHCH21] Lai, Cheng-Chien, Wei-Hsin Huang, Betty Chia-Chen Chang and Lee-Ching Hwang: *Development of machine learning models for prediction of smoking cessation outcome.* International journal of environmental research and public health, 18(5):2584, 2021.

[LLQ+20] Liu, Na, Xiaomei Li, Ershi Qi, Man Xu, Ling Li and Bo Gao: *A novel ensemble learning paradigm for medical diagnosis with imbalanced data.* IEEE Access, 8:171263–171280, 2020.

[MMPM23] Mohapatra, Subasish, Sushree Maneesha, Prashanta Kumar Patra and Subhadarshini Mohanty: *Heart Diseases Prediction based on Stacking Classifiers Model.* Procedia Computer Science, 218:1621–1630, 2023.

[NA22] Nasiri, Hamid and Seyed Ali Alavi: *A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images.* Computational intelligence and neuroscience, 2022(1):4694567, 2022.

[NB23] Nguyen, Hung Viet and Haewon Byeon: *Prediction of Parkinson's disease depression using LIME-based stacking ensemble model.* Mathematics, 11(3):708, 2023.

[NRP+22] Nikam, Mihir, Ameya Ranade, Rushil Patel, Prachi Dalvi and Aarti Karande: *Explainable Approach for Species Identification using LIME.* In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE, 2022.

[OCP+20] Ortis, Alessandro, Pasquale Caponnetto, Riccardo Polosa, Salvatore Urso and Sebastiano Battiato: *A report on smoking detection and quitting technologies.* International journal of environmental research and public health, 17(7):2614, 2020.

[Sho21] Shorewala, Vardhan: *Early detection of coronary heart disease using ensemble techniques.* Informatics in Medicine Unlocked, 26:100655, 2021.

[SPN23] Sonawani, Shilpa, Kailas Patil and Prabhu Natarajan: *Biomedical signal processing for health monitoring applications: a review.* International Journal of Applied Systemic Studies, 10(1):44–69, 2023.

[SS20] Singh, Namrata and Pradeep Singh: *Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus.* Biocybernetics and Biomedical Engineering, 40(1):1–22, 2020.

[SVA+22] Swapna, Mudrakola, Uma Maheswari Viswanadhula, Rajanikanth Aluvalu, Vijayakumar Vardharajan and Ketan Kotecha: *Bio-signals in medical applications and challenges using artificial intelligence.* Journal of Sensor and Actuator Networks, 11(1):17, 2022.

[VMH+20] Volk, Robert J., Tito R. Mendoza, Diana S. Hoover, Shawn P.E. Nishi, Noah J. Choi and Therese B. Bevers: *Reliability of self-reported smoking history and its implications for lung cancer screening.* Preventive Medicine Reports, 17:101037, 2020.

[VSM24] Vimbi, Viswan, Noushath Shaffi and Mufti Mahmud: *Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection.* Brain Informatics, 11(1):10, 2024.

[Wor23] World Health Organization: *Tobacco*, 2023. Accessed: 2023-08-26.

[YKS20] Yang, Xiaoyan, Matloob Khushi and Kamran Shaukat: *Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction.* In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2020.

[ZSL21] Zheng, Huilin, Syed Waseem Abbas Sherazi and Jong Yun Lee: *A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data.* IEEE Access, 9:113692–113704, 2021.

[ZTW$^+$24] Zhang, Xiaoshuai, Chuanping Tang, Shuohuan Wang, Wei Liu, Wangxuan Yang, Di Wang, Qinghuan Wang and Fang Tang: *A stacking ensemble model for predicting the occurrence of carotid atherosclerosis.* Frontiers in Endocrinology, 15:1390352, 2024.

[23] , and : *Data preprocessing techniques in machine learning.* , 1(2), 2023.