iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58615

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Predictive Modelling to Understand and Combat Smoking Using Machine Learning: A Review

Irfan Aijaz[1], Dr. Gurinder Kaur Sodhi[2]
[1]*M. Tech Scholar, Department of ECE Engineering, Desh Bhagat University, Mandi Gobindgarh Punjab, India*
[2]*Assistant Professor, Department of ECE Engineering, Desh Bhagat University, Mandi Gobindgarh Punjab, India*

Abstract: ***Smoking remains a worldwide public health issue, with significant personal consequences well-being and society as a whole. Traditional approaches to understanding and combatting smoking have their limitations, and in recent years, machine learning has become a viable instrument to tackle this problem. This review article provides a comprehensive overview of predictive modeling to understand and combat smoking using machine learning. We delve into the diverse data sources and preprocessing techniques, feature engineering approaches, and machine learning models employed in the context of smoking prediction. The review categorizes studies into smoking initiation and smoking cessation prediction, shedding light on the methodologies, results, and challenges in each domain. Furthermore, we explore the real-world applications of predictive modeling in smoking control, emphasizing their impact on public health policy and awareness campaigns. Ethical considerations and challenges related to bias, privacy, and model interpretability are also discussed. The paper concludes by suggesting future research directions and emphasizing the crucial role of machine learning in comprehensively addressing the smoking epidemic..***

Keywords: Preventive healthcare, Predictive Modelling, Personalized healthcare, Multimedia applications.

I. INTRODUCTION

Smoking, a deeply ingrained yet highly detrimental habit, has long remained a global public health challenge. The consequences of smoking are both personal and societal, with severe health implications for individuals and substantial economic and healthcare burdens for societies. [1]. Despite decades of concerted efforts to reduce smoking prevalence, this addictive behavior continues to persist and pose a significant threat to public well-being. In the quest to understand and combat smoking, traditional approaches have played a vital role. However, these approaches have faced inherent limitations, making it increasingly imperative to explore innovative methodologies[2].

This review paper embarks on a comprehensive journey through the realm of predictive modeling to understand and combat smoking using the capability of artificial intelligence. Artificial intelligence's machine learning branch has acquired prominence for its ability to extract meaningful insights from large and complex datasets. By harnessing this technology, researchers and public health practitioners have sought to gain a deeper understanding of smoking behaviors, predict smoking initiation, and develop effective strategies for smoking cessation[3].

The significance of machine learning in this instance is its ability to reveal hidden trends and relationships within data that were previously inaccessible through traditional methods. It offers the potential for precision, individualized interventions that can transform the landscape of smoking control and public health policy[4]. With machine learning, predictive models can not only identify at-risk individuals but also tailor interventions to their unique needs, making smoking prevention and cessation efforts more targeted and effective.

The objectives of this review paper are twofold. First, we aim to provide an exhaustive overview of the existing literature on predictive modeling for smoking control using machine learning[5]. By exploring the various data sources, preprocessing techniques, feature engineering approaches, and machine learning models employed in this domain, we offer readers a comprehensive understanding of the methodologies that underpin this evolving field.

Second, we categorize the literature into two key domains: predictive modeling for smoking initiation and predictive modeling for smoking cessation[6].

In doing so, we highlight the unique challenges and opportunities present in each context. Additionally, we examine the practical uses of algorithmic learning in smoking cessation, investigating their impact on public health policy, smoking cessation programs, and awareness campaigns[7].

To navigate this complex landscape, it is essential to address the ethical implications and challenges that arise when applying machine learning to public health issues. Bias, privacy concerns, and model interpretability are just a few of the ethical considerations that warrant thoughtful discussion[8].

Ultimately, as we conclude this review paper, we offer insights into potential future research directions, highlighting emerging technologies and methodologies that hold the promise of further advancing our understanding and combatting of smoking. With a growing body of research and a deepening appreciation for the capabilities of machine learning, we stand at the precipice of a new era in smoking control and public health[9].

II. DEVELOPMENT OF MACHINE LEARNING-BASED MULTI-CLASS CLASSIFICATION MODELS

A. *Feature Selection Analysis*

Feature selection stands as a cornerstone in the quest to enhance the predicted reliability and efficacy of our categorization methods. We have used the wrapper-based Sequential Backward Feature Selection (SBFS) strategy in our work, which is essential for lowering the size of a high-dimensional feature space while preserving the most relevant and informative features. The core objectives of feature selection encompass:

1) *Identifying Relevant Features:* The central aim of the feature selection process is to identify and retain those features that significantly contribute to improving the predictive performance of our classification models.
2) *Addressing Dimensionality Challenges:* High-dimensional feature spaces often present the challenge of the "curse of dimensionality," particularly when dealing with limited data. Feature selection serves as a potent remedy to mitigate this issue.
3) *Eliminating Noise and Irrelevance:* The feature selection process is instrumental in removing noise and irrelevant features from the dataset, thereby significantly enhancing the overall accuracy and interpretability of our classification models.
4) *Reducing Computational Complexity:* By carefully choosing a group of characteristics that are most discriminative and informative, feature selection significantly contributes to the reduction of computational time and resources required for the training and prediction phases of our models.

In our study, the implementation of SBFS made effective use of the mlxtend Python library [47]. This systematic and efficient approach enabled us to select the optimal set of features, which in turn greatly impacted the performance of our classification models.

B. *Development of Classification Models*

The development of our classification models represents a core component of our endeavor to accurately detect smoking activity. These models are designed to assign class labels to new input feature vectors and have been rigorously trained using a carefully labeled dataset. In our pursuit of the most effective predictive models, we have scrutinized the following classification techniques for their predictive strength:

1) *Logistic Regression (LR)*

Logistic Regression, a widely employed linear classification model, is celebrated for its simplicity and interpretability. It is particularly well-suited for both binary and multi-class classification tasks, making it a versatile and valuable choice [12, 67]. A popular statistical modeling method is logistic regression, which links the likelihood of an event to a number of possible predictors. LR is a method that is becoming more and more common for modeling the likelihood of discrete situations. When used correctly, LR may provide highly strong insights into whether traits have a greater or lesser chance of predicting an event's outcome for a group of interest [6]. LR describes mathematical frameworks in which the outcome inconsistent or dependent variable, becomes categories as opposed to indefinite. Changes in the meanings of the continuously or dichotomous independent variables that are located on the right part of the expression are "mapped" or "translated" by the logistic function to represent an increase or decrease in the chance of the event that the dependent, or left-side, component is modeling. When using a logistic regression model, . Often, logistic regression is a better choice [7]. Class rates are easily estimated by logistic regression. Superior than J48, zeroR, and naïve bayes [8].

2) *K-Nearest Neighbors (KNN)*

A informal sorting method called K-Nearest Neighbors assigns labeling for classes to data points according to the general group of their the k closest neighbors.. Notable for its simplicity and ease of interpretation, KNN has proven to be an invaluable asset in our classification framework [4, 10].

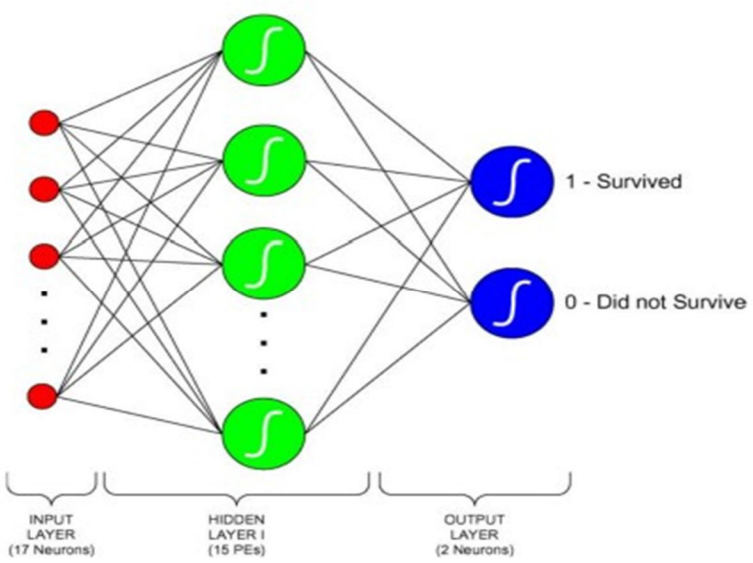


Figure 1 KNN method for smoke detection

3) Adaptive Boosting (AdaBoost)

Adaptive Boosting, an ensemble learning method, excels at combining multiple weak learners to create a strong, accurate classifier. Its strength lies in its capacity to adapt and improve performance iteratively, rendering it an indispensable part of our classification ensemble [18, 19].

4) Random Forest (RF)

The Random Forest technique, another Using a team-learning technique, a number of decision trees are built, and predictions from them are combined to reach robust and reliable results. Its resistance to overfitting and consistently high accuracy make it an essential component of our classification framework [9].

5) Support Vector Machines (SVM)

Noted for their resilience and efficiency, supporting vector machines are a potent form of system that looks for an orbit with the largest possible margin of classes. They excel in both binary and multi-class classification tasks and have become a cornerstone of our classification ensemble [13].

6) Decision Tree Classifier (DT)

The Decision Tree Classifier, a simple yet interpretable model, recursively splits data into subsets based on feature values. It generates decision trees that make predictions by navigating through a series of binary decisions based on feature values, offering transparency and efficiency in our classification framework [12].

These diverse classification models are instrumental in building a robust system for detecting smoking activity. Each model contributes its unique strengths and advantages to the overall framework. The development process is further enriched through meticulous evaluation, which includes both in-sample and out-of-sample validation, as well as hyper-parameter tuning to ensure the optimal performance of the models. A decision tree (DT) algorithm displays the learning function as a set of if_then_else rules or as a decision tree. DTs have several shortcomings while being the easiest functions to comprehend and interpret [9]. A DT's production requires a lot of processing power. Algorithms used to generate DT are unstable. Unusual quality choices inside the tree at each node may arise from little modifications in the training data. Because attribute selection can impact each and every sub tree, the total impact on categorization is substantial. Therefore, the resulting regulations' complexity may rise. Additionally, because over-fitting of training data occurs due to the DT building approach's predominant usage of greedy heuristics, predicted future accuracy is negatively impacted [10].

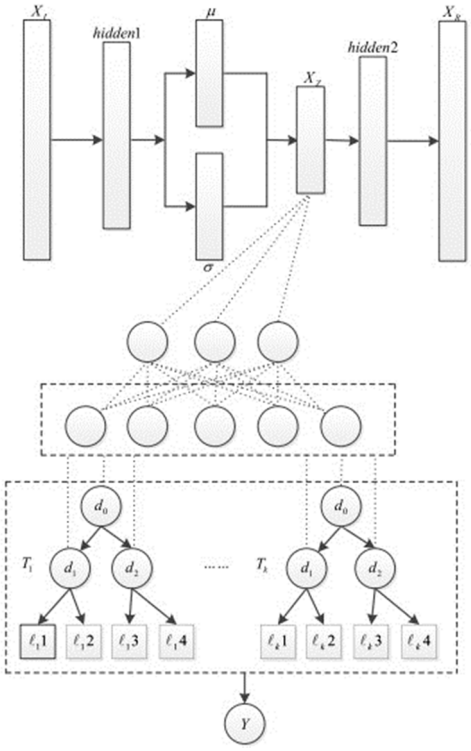


Figure 2 Framework of the hybrid model

III. IN-SAMPLE AND OUT-OF-SAMPLE VALIDATION

A. *Classification Approaches: In-Sample and Out-of-Sample Validation*

The robustness and reliability Effective multi-class classification frameworks using machine learning to identify smoking behavior depend on how well their prediction power is assessed. This section covers the adjustments of network hyper-parameters and the crucial components of both in-sample and out-of-sample validity.. These validation methods serve distinct purposes, and hyper-parameter tuning contributes to optimizing the performance and efficiency of the classification models.

B. *In-Sample Validation*

In-sample validation is the first lens through which the established classifications networks' ability to forecast is assessed. This approach involves utilizing the training data to make predictions using the models that have been meticulously constructed. In essence, it is an assessment conducted within the familiar confines of the training dataset. The primary goal of in-sample validation is to establish a reference point or a maximum value for the models' prediction accuracy.

During in-sample validation, the classification models are rigorously examined and tested using the very data on which they were trained. This process allows for the measurement of the models' capacity to fit the training data accurately. The assessment includes scrutinizing the models' ability to recognize patterns, relationships, and relevant features within the data. Such an approach offers an initial glimpse into the models' performance, serving as a benchmark for comparison with subsequent validation methods.

C. *Out-of-Sample Validation*

In stark contrast to in-sample validation, out-of-sample validation presents a more stringent and revealing assessment of the classification models. This validation method employs a distinct dataset—the test data—that has not been encountered by the models during the training phase. The fundamental objective of out-of-sample validation is to determine the models' generalizability, or their ability to perform effectively on entirely new, unseen data. Out-of-sample validation is akin to a litmus test for the models' practical utility. It simulates real-world scenarios where the models encounter novel cases or individuals. By introducing previously unseen data, this validation method effectively gauges the models' ability to extrapolate from the training data and apply their learned patterns to new situations. The outcomes of out-of-sample validation provide valuable insights into how well the classification models are likely to perform in real-world applications and decision-making scenarios.

D. Model Hyper-Parameter Tuning

The quest for optimizing the performance and efficiency of the classification models extends beyond validation and into the realm of hyper-parameter tuning. In this endeavor, the process of grid search, a hyper-parameter optimization technique, is employed in conjunction with 5- Fold cross-validation. Finding the best additional parameters for identifying models is the main goal of this refinement procedure.

Among the essential elements that control actions and performance of machine learning models. They include factors like learning rates, regularization strengths, and other model-specific settings. The selection of appropriate hyper-parameters is instrumental in fine-tuning the models for optimal performance. The combination of grid search and cross-validation offers a systematic and thorough approach to this task.

Grid search systematically explores various hyper-parameter combinations within predefined ranges. The efficacy of each combination is assessed through cross-validation, a process that partitions the data into subsets for training and testing, subsequently validating the model's performance. This iterative exploration helps identify the hyper-parameters that yield the best results, ultimately enhancing the models' accuracy and generalizability.

IV. EVALUATION METRICS

In the quest for effective smoking activity detection and classification, it is imperative to employ robust evaluation metrics that provide a comprehensive understanding of the models' performance. In this section, we delve into the evaluation metrics, with a particular emphasis on the receiver operative parameter curve's (ROC) region underneath the curve (AUC). Furthermore, we explore the critical issue of determining the optimal window size for feature extraction from raw sensor data, which plays a pivotal role in achieving accurate and reliable smoking activity detection.

A. Evaluation Metrics: Area under the ROC Curve (AUC)

Evaluation metrics are essential in assessing the quality and effectiveness of classification models. Among the plethora of metrics available, In our research, the area under the Receiver Operating Characteristic, or ROC, curve Report Phrase (AUC) serves as a common benchmark. A graphical representation of the conflict amongst a strategy sensitization (true optimistic rate) and breadth (true neg ratio) is called a ROC curve. across varying classification thresholds. AUC, in particular, quantifies the overall discriminatory power of a model and its ability to distinguish between different classes. It is an integral component of the evaluation process for classification models in our study. AUC values vary from 0 to 1. A model's ability to perform is comparable to erratic guessing at a value of 0.5, while a perfect classifiers is indicated by a value of 1. The utilization of AUC as a key evaluation metric allows us to not only assess the overall performance of the classification models but also make informed comparisons between different models. A higher AUC indicates a model's superior ability to discriminate between classes, reflecting its higher predictive accuracy.

B. Optimal Window Size Analysis

The choice of window size for feature extraction from raw sensor data is a critical consideration in our pursuit of accurate smoking activity detection. To gain insights into the impact of varying window sizes, we performed a detailed comparative analysis involving different window sizes, specifically 1 second, 3 seconds, and 5 seconds. This analysis is vital in the selection of the most appropriate window size for feature extraction, a decision that directly influences the accuracy and reliability of smoking activity detection.

The selection of the optimal window size hinges on several factors, including the nature of the sensor data, the type of activities to be detected, and the trade-off between time resolution and feature informativeness. Our analysis involves a thorough examination of the outcomes and performance of the classification models when fed with feature vectors generated using different window sizes.

A key aspect of this analysis is the consideration of the temporal characteristics of smoking activities and the specific hand movements associated with them. Things like snacking, drinking, jogging, smoking, and traveling often involve significant hand movements. Therefore, the window size must be chosen to effectively capture these movements and translate them into informative features for classification.

In our study, a systematic approach was adopted to experiment including windows of one, three, and five seconds, among other sizes. A unique harmony between temporal clarity and the quantity of data accessible for feature extraction is provided by every single window size. Each window's raw data is painstakingly digested to get the matching feature vector, which includes a variety of evocative, in the time domain, and frequency-domain properties.

The comparative analysis of these window sizes is integral to our quest for optimal feature extraction. The ultimate aim is to identify the window size that maximizes the accuracy of smoking activity detection while ensuring the efficiency of the system. This determination is informed by the trade-offs between time resolution, computational efficiency, and the need to capture relevant features that reflect smoking behaviors accurately.

V. CONCLUSIONS

In this comprehensive review paper, we have explored the multifaceted landscape of predictive modeling in the context of understanding and combatting smoking addiction using machine learning. Smoking remains An danger to global healthcare, and the incorporation of sophisticated technology and predictive modeling techniques holds promise for promoting smoking cessation and improving public health outcomes.

Our journey through the realm of smoking cessation efforts commenced with an investigation of diverse data sources and preprocessing techniques. These include the utilization of wearable sensor devices, smartphone applications, and electronic health records, all of which offer rich sources of data for understanding smoking behaviors. The preprocessing steps encompassed data cleaning, feature extraction, and the integration of clinical data to prepare the ground for predictive modeling.

We delved into the various methodologies employed to encourage smoking cessation. Pharmacological treatments, such as nicotine replacement therapy (NRT), and non-pharmacological treatments, including behavioral therapy, counseling, and the use of smartphone applications, have emerged as effective strategies. A crucial finding emphasized the significance of combo therapy, which combines coaching and NRT, in promoting successful smoking cessation.

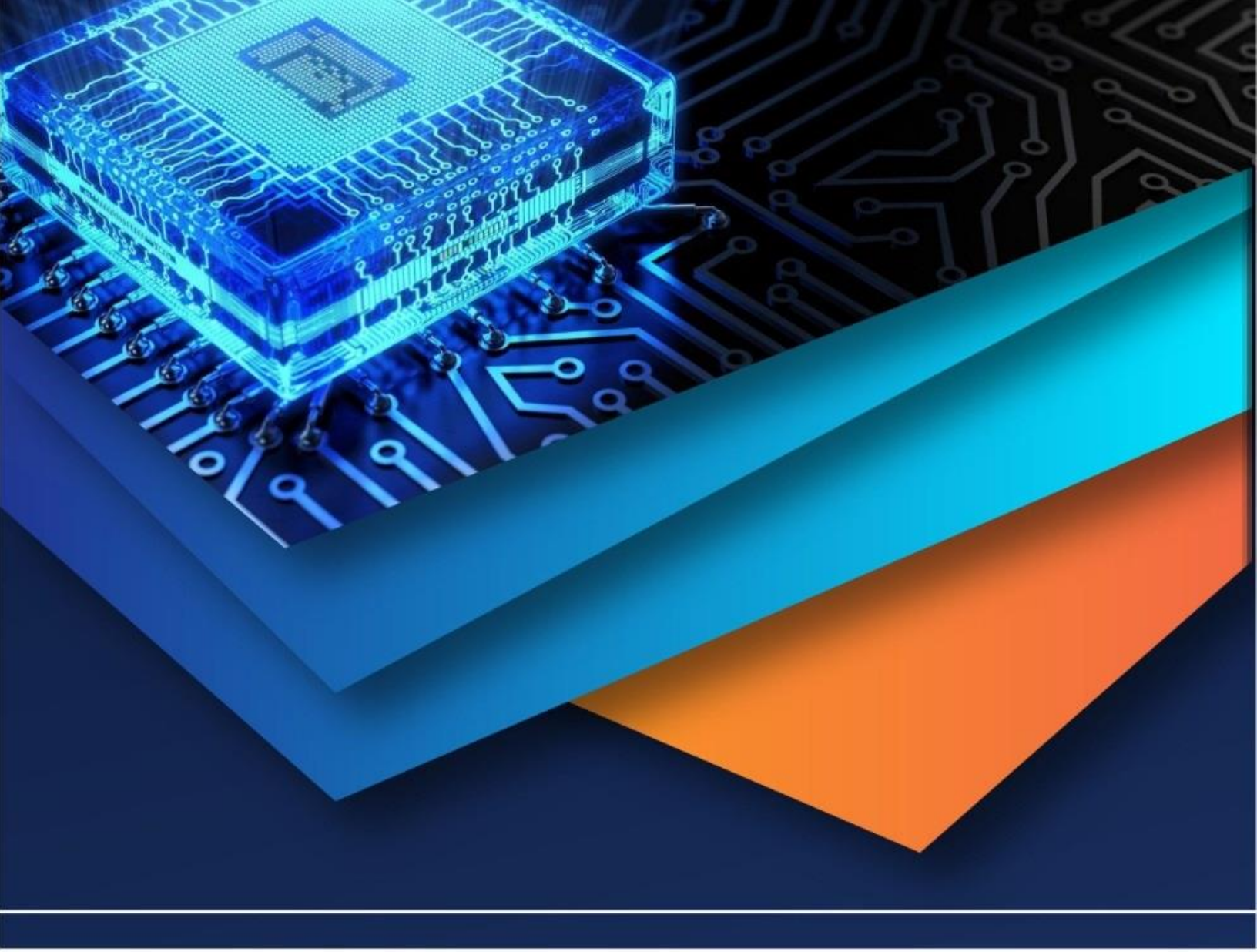
Furthermore, our exploration ventured into the realm Developing multi-class classification models for vaping identification based on machine learning methods. We discussed in detail how machine learning models for classification are developed, how features are chosen, and how assessment metrics—most notably the area under the curve of the Receiver Operating Characteristic (ROC) (AUC)—are used to gauge how well a model is doing.

An examination of the ideal window dimensions was conducted to determine the window size that yields the highest accuracy for feature extraction from raw sensor data. This in-depth investigation was a critical aspect of our quest for reliable smoking activity detection.

REFERENCES

[1] Abroms LC, Lee Westmaas J, Bontemps-Jones J, Ramani R, Mellerson J (2013) A content analysis of popular smartphone apps for smoking cessation. Am J Prev Med 45(6):732–736
[2] Adibi S (2015) Mobile health a technology road map, 5th edn. Springer International Publishing
[3] Akash K, Hu W-L, Jain N, Reid T (2018) A classification model for sensing human trust in machines using EEG and GSR. ACM Trans Interact IntellSyst 8(4):1–20
[4] Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3): 175–185
[5] Al-Ubaydli O, List JA, LoRe D, Suskind D (2017) Scaling for economists: lessons from the non-adherence problem in the medical literature. Econ Perspect 31(4):125–144
[6] Aly M (2005) Survey on multiclass classification methodsAnguita D, Ghio A, Oneto L, Ridella S (2012)
[7] In-sample and out-of-sample model selection and error estimation for support vector machines. IEEE Trans Neural Networks Learn Syst 23(9):1390–1406
[8] Atallah L, Lo B, King R, Yang G-Z (Aug. 2011) Sensor positioning for activity recognition using wearable accelerometers. IEEE Trans Biomed Circuits Syst 5(4):320–329
[9] Breiman L (2001) Random forests. Mach Learn 45(1):5–32
[10] Burba F, Ferraty F, Vieu P (2009) K -nearest neighbour method in functional nonparametric regression. J Nonparametr Stat 21(4):453–469
[11] Chatterjee S, Price A (2009) Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Informatics Assoc 16(2):171–178
[12] Cox DR (1958) The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) 20. WileyRoyal Statistical Society:215–242
[13] Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines : and other kernel-based learning methods. Cambridge University Press
[14] Erdaş B, Atasoy I, Açici K, Oğul H (2016) Integrating features for accelerometer-based activity recognition. Proc Comput Sci 58:522–527
[15] Eyobu OS, Kim YW, Cha D, Han DS (2018) A real-time sleeping position recognition system using IMU sensor motion data. In: 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, vol 2018, pp 1–2
[16] Fogg BJ (1999) Persuasive technologies. Commun ACM 42(5):26–29
[17] Formagini TDB, Ervilha RR, Machado NM, de Andrade BABB, Gomide HP, Ronzani TM (2017) A review of smartphone apps for smoking cessation available in Portuguese. Cad Saude Publica 33(2)
[18] Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J ComputSyst Sci 55(1):119–139
[19] Freund Y, Schapire RE (1999) A short introduction to boosting. J Japanese Soc ArtifIntell 14(5):771–780

[20] Gao Z, Xuan HZ, Zhang H, Wan S, Choo KKR (2019) Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet Things J 6(6):9280–9293
[21] Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
[22] Haskins BL, Lesperance D, Gibbons P, Boudreaux ED (2017) A systematic review of smartphone applications for smoking cessation. Translational Behavioral Medicine 7(2) Springer New York LLC: 292–299
[23] Heydari G et al (2014) A comparative study on tobacco cessation methods: A quantitative systematic review. International Journal of Preventive Medicine 5(6) Isfahan University of Medical Sciences:673–678
[24] Heydari G et al (2015) Assessment of different quit smoking methods selected by patients in tobacco cessation centers in Iran. Int. J. Prev. Med 2015
[25] Hoeppner BB, Hoeppner SS, Seaboyer L, Schick MR, Wu GWY, Bergman BG, Kelly JF (2016) How smart are smartphone apps for smoking cessation? A content analysis. Nicotine Tob Res 18(5):1025–1031
[26] . Huang WH, Hsu HY, Chang BCC, Chang FC (2018) Factors correlated with success rate of outpatient smoking cessation services in Taiwan. Int J Environ Res Public Health 15(6)
[27] Jain R, Majumder P, Gupta T, Bandiera SM (2013) Pharmacological Intervention of Nicotine Dependence. Biomed Res Int 2013
[28] Jha P, Peto R (2014) Global effects of smoking, of quitting, and of taxing tobacco. N Engl J Med 370(1): 60–68
[29] Kim E, Lee J, Shin H, Yang H, Cho S, Nam SK, Song Y, Yoon JA, Kim JI (2019) Champion-challenger analysis for credit card fraud detection: hybrid ensemble and deep learning. Expert Syst Appl 128:214–224
[30] Kosse NM, Brands K, Bauer JM, Hortobagyi T, Lamoth CJC (2013) Sensor technologies aiming at fall prevention in institutionalized old adults: A synthesis of current knowledge. International Journal of Medical Informatics 82(9) Elsevier:743–752
[31] Luna-Perejon F, Malwade S, Styliadis C, Civit J, Cascado-Caballero D, Konstantinidis E, Abdul SS, Bamidis PD, Civit A, Li YC(J) (2019) Evaluation of user satisfaction and usability of a mobile app for smoking cessation. Comput Methods Prog Biomed 182:105042
[32] McClure JB, Hartzler AL, Catz SL (2016) Design considerations for smoking cessation apps: feedback from nicotine dependence treatment providers and smokers. JMIR mHealth uHealth 4(1):e17
[33] Méndez D, Tam J, Giovino GA, Tsodikov A, Warner KE (2016) Has Smoking Cessation Increased? An Examination of the US Adult Smoking Cessation Rate 1990–2014. Nicotine Tob Res:ntw239
[34] Messer K, Trinidad DR, Al-Delaimy WK, Pierce JP (Feb. 2008) Smoking cessation rates in the United States: a comparison of young adult and older smokers. Am J Public Health 98(2):317–322
[35] Miao F, He Y, Liu J, Li Y, Ayoola I (2015) Identifying typical physical activity on smartphone with varying positions and orientations. Biomed. Eng. Online 14(32)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY