2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

Decision Tree Model of Smoking Behaviour

Maryam Abo-Tabik
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University
Manchester, United Kingdom
maryam.a.abo-tabik@stu.mmu.ac.uk

Nicholas Costen
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University
Manchester, United Kingdom
N.Costen@mmu.ac.uk

John Darby
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University
Manchester, United Kingdom
J.Darby@mmu.ac.uk

Yael Benn
Department of Psychology
Manchester Metropolitan University
Manchester, United Kingdom
Y.Benn@mmu.ac.uk

***Abstract*—Smoking is considered the cause of many health problems. While most smokers wish to quit smoking, many relapse. In order to support an efficient and timely delivery of intervention for those wishing to quit smoking, it is important to be able to model the smoker's behaviour. This research describes the creation of a combined Control Theory and Decision Tree Model that can learn the smoker's daily routine and predict smoking events. The model structure combines a Control Theory model of smoking with a Bagged Decision Tree classifier to adapt to individual differences between smokers, and predict smoking actions based on internal stressors (nicotine level, withdrawal, and time since the last dose) and external stressors (e.g. location, environment, etc.). The designed model has 91.075% overall accuracy of classification rate and the error rate of forecasting the nicotine effect using the designed model is also low (MSE=0.048771, RMSE=0.216324, and NRMSE=0.153946) for regular days and (MSE=0.048804, RMSE=0.216637, and NRMSE=0.195929).**

***Index Terms*—smoker's behaviour, addictive behaviour, machine learning, Decision Tree, Bagged Decision Tree, Control Theory.**

I. Introduction

Smoking is considered one of the leading causes of deaths internationally. According to a recent NHS report [1] in 2016, smoking caused the death of about 77,900 people in England alone. The report further states that smoking is not only harmful to the smokers, but many diseases might be caused by the exposure to passive smoking, especially affecting children who are particularly vulnerable to the effects of passive smoking. This makes reducing cigarette smoking a public health priority.

Actions (including smoking) can be seen as being motivated by the need of the human system to maintain stability, over a range of time-scales, in the face of a changing environment. This motivation can appear in the form of internal feelings such as sadness, or external need such as maintaining nicotine level [2]. Closed-loop control model is a common instrumental modelling method that seeks to maintain stability. It employs the feedback principle, useing the output data from the model (feedback signal) as an input to modify the model's actions, and hence maintain stability [3]. However, modelling addictive behaviour as a closed loop control model is a challenging task. It requires understanding the complexity of humans, as well as determining what elements should be counted to model the addictive behaviour. Moreover, when modelling the addictive behaviour, the goal state represents the fact that the system seeks to obtain a steady state (natural state), rather than to imply that there exists a single fixed value, as is often the case in system engineering [4].

Opponent process theory is claimed to be an essential method that can be used to model a person's emotional state [5]. Solomon [6] described addictive behaviour using the opponent process theory. Within this model, an addict experiences pleasure as soon as a drug is supplied, which is followed by slowly accumulated withdrawal symptoms. As such, during the initial stages of addiction, the pleasure level is high and is accompanied by a low level of withdrawal symptoms. However, as time goes by, the withdrawal symptoms increase leading to a decrease in pleasure caused by using the drug, potentially resulting in a higher quantity of the drug being consumed [4].

Bobashev et al. [7] modelled the behaviour of smokers and employed the opponent process scheme of control theory. The model did not present any complex neurobiological process, only providing a mathematical model with a cascading feedback loop, aimed at presenting the scientific narrative of the opponent process as shown in Fig. 1.

The model equations were developed with phenomenological interpretation in mind, and no real biological process was modelled. A set of continuous functions were used feed into the cascading functions. The system equations involve five interlinked processes,

$$ProcessA : \frac{dY_1}{dt} = e^{-\alpha t} - b_1 Y_1 \tag{1}$$

978-1-7281-4034-6/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00311

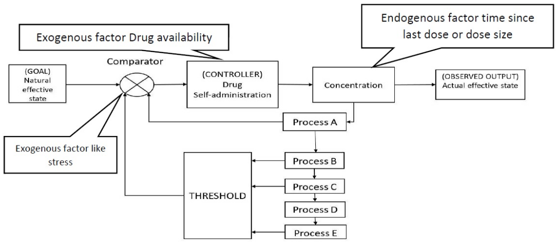

Exogenous factor Drug availability
Endogenous factor time since last dose or dose size
Comparator
(GOAL) Natural effective state
(CONTROLLER) Drug Self-administration
Concentration
(OBSERVED OUTPUT) Actual effective state
Process A
Exogenous factor like stress
THRESHOLD
Process B
Process C
Process D
Process E

Fig. 1: Control theory model of smoking based on [7].

$$ProcessB : \frac{dY_2}{dt} = a_1Y_1 - b_2Y_2 \tag{2}$$

$$ProcessC : \frac{dY_3}{dt} = a_2Y_2 - b_3Y_3 \tag{3}$$

$$ProcessD : \frac{dY_4}{dt} = a_3Y_3 - b_4Y_4 \tag{4}$$

$$ProcessE : \frac{dY_5}{dt} = a_4Y_4 - b_5Y_5 \tag{5}$$

where a, b and α are scaling coefficients, and all the Y_i initial values are equal set to zero. Each equation presents a weighted integration of the previous one, causing the processes to lengthen sucessively. Y_1 represented the effect of nicotine level and is modelled with a pharmacokinetic equation. Y_2 represents the toxicity level and how the body processes the drug. Y_3 is the daily smoking habit. Y_5 is a longer scaling habit, which is scaled in years (rather than minutes/ hours/ days). While the process Y_4 has not been interpreted, it has been used to add scaling period between Y_3 and Y_5, which results in a slow change in process Y_5. To simulate smoking behaviour, a threshold value was defined to prompt self-administration. The threshold

$$T = \frac{(\beta_3Y_3 + \beta_5Y_5)}{(1 + \beta_2Y2)} \tag{6}$$

has calibration coefficients β_i, and to avoid division by zero one is added to the denominator of the equation. The threshold value is changed based on external stressors to initiate cigarette use

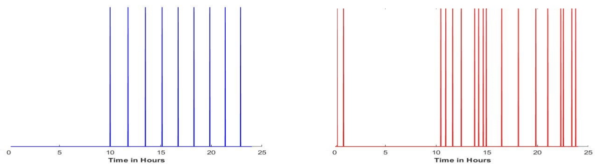
$$T = T + stress. \tag{7}$$

The research also modelled the withdrawal and craving processes; these p rocesses begin immediately following the initial nicotine use and grow over time

$$W = \frac{d_3Y_3(T - Y_1)}{(Y_{0w} + Y_1)} \tag{8}$$

$$C = \frac{d_5Y_5(T - Y_1)}{(Y_{0c} + Y_1)} \tag{9}$$

where d_3, d_5, Y_{0w} and Y_{0c} are calibration coefficients. This control theory model was able to simulate the changes in smoking behaviour over time. However, the system was not able to present real-life behaviour, and could not capture individual differences between smokers' daily habits. Fig. 2 presents the differences between the smoking behaviour as presented using the simulated control theory model Fig. 2a and real-life data collected from a participant shown in Fig. 2b.



(a) Simulated smoking events

(b) Real smoking events

Fig. 2: Smoking frequency; each peak represents a smoking event (a) a simulated smoking behaviour generated by the control theory model [7] , and (b) real smoking behaviour on a randomly-selected day from our collected data.

Studies show that modelling smoking behaviour is essential, it can improve the intervention process in the way of helping smokers in their most needed time [8]. While most of the known approaches try to find a relationship between some clues (e.g., withdrawal, stress, place, and the presence of other smokers) and urge to smoke. Most of these studies rely on participants self-reporting these indicators, as the results indicated that these predictors provide a high degree of possibility for predicting potential smoking events or relapse in quitting period. However, Self-reporting as a method can be inaccurate as it is sensitive to self-biased errors [9]. Another research [10] investigates the possibility of using hidden Markov models to set patterns for the timing and places that the smokers are most likely to smoke, and then use these patterns for better delivery of the support messages. The paper did not report any analytical result that is related to Hidden Markov models, except the positive feedback from the participants who used their mobile application.

As such, the current research aims to develop a machine learning model, which when combined with a control theory model of smoking, will be able to adapt to the smoker's unique behaviour and predict future smoking events. The Bobashev et al. [7] model was chosen due to its ability to capture the nicotine effect using the pharmacokinetic equation. Here, we describe the implementation of this control theory model of smoking that is expanded to incorporate other factors affecting smokers' smoking behaviour (e.g., location and activity).

II. Decision Tree for classifying unimodal tabular data

Many classification problems have a large dataset containing complex information, including potential labelling inaccuracies. A decision tree is considered to be an efficient machine learning classifiers for such problems [11]. An early version of the regression tree is the classification and regression tree CART[12]; it recursively divided the dataset based on the selected features using Least Squared Deviation (LSD) as its impurity function [13],

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i (y_i - \bar{y}_i(t))^2 \quad (10)$$

$$\bar{y}(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i y_i \quad (11)$$

$$N_w(t) = \sum_{i \in t} w_i f_i \quad (12)$$

where $N_w(t)$ is the weighted number of samples in node t, w_i is the calculated weight value for each i and f_i is the recorded response. y_i is the response and $\bar{y}_i$ is the value of the mean response. t The splitting process is performed using

$$Q(s,t) = R(t) - R(t_L) - R(t_R) \quad (13)$$

where t_L is the left child and t_R is the right child of the node t.

Many classifications enhanced their models by training their dataset using several classifiers, and the results are then combined using a voting process, this general method being called an ensemble classifier [14]. The ensemble has also been used with decision trees, mainly in two approaches; either Bagging [15], or boosting [16] algorithms. Bagging (or bootstrap aggregating) is applied to decision trees by generating multiple versions of decision trees during the training process and using a plurality vote between them to predict the class. The idea is to create several subsets from the dataset, with each subset training its own decision tree, and then combine the result from several trained models in order to reach a more reliable predictor and reduce the variance of classification [17, 18]. Boosting is the use of iterative re-training, so as to create the ensemble sequentially, where at each step the later trained classifier is learning from the previous errors generated by the earlier classifiers, by increasing the weight as the training progresses [19]. While boosting classifiers increases the accuracy of the trained model over bagging, in return it increases the chance of overfitting; another drawback for boosting is that it is very slow, and it is sensitive to noise [20].

Another form of the ensembles decision tree is the random forest; this model is efficient because it reduces the overfitting problem [21]. Random forest randomly selects subset samples from the training set (in-bagging) and use them to generate multiple versions of the decision tree. The rest of the samples (out-bagging) will be used in cross-validation to estimate how well the classifier works. The generated error from the validation process is called out-of-bag (OOB) error. Random forest is automatically produced without any pruning, and each node splits using a predefined number of features. The forest grows up to a set limit of the number of trees. Random forest generates trees with low bias and high variance. The classification output is calculated by averaging the class assignment probability generated by all the trees; the probability of the class is calculated using all the produced trees [22].

III. Data collection and processing

There is currently no published dataset that can fit the needs of our research. Moreover, to create a data set that can be employed in modelling smoking behaviour, several steps were followed. A mobile application was used to collect signals from mobile sensors (e.g., movement and environment) for approximately two weeks, while users reported their smoking events. Three types of events occurr in the dataset, which are labelled as smoking (1), not smoking (2) and app-off (0) events. The later was labelled as app-off due to gaps in the dataset (i.e. the participant's mobile phone was off). Table I shows the frequency of events for each of the four participants. One problem that can be seen is that the classes are unbalanced, as the number of smoking events is much lower than the number of non-smoking events. Overall, there are 1440 data samples per day (one sample per minute), while the reported smoking events are less than 15 per day, and the rest are either not smoking or app-off events. To overcome this problem the model is targeting at the smoking period, not at the per-minute smoking event. Instead, the data labeling changed to include a 10-minute window, hence reducing the ratio of smoking to non-smoking events. Table II shows the frequency of events for each of the four participants after applying the change.

TABLE I: The number of labels in each of the three labelling categories.

	App off	**Smoking**	**Not smoking**
Participant 1	451	201	18068
Participant 2	6307	64	12349
Participant 3	3997	66	14657
Participant 4	15514	82	3124

TABLE II: The number of labels in each of the three labelling categories after applying a 10 minute smoking window.

	App off	**Smoking**	**Not smoking**
Participant 1	451	1960	16308
Participant 2	6217	630	11872
Participant 3	3997	650	14072
Participant 4	15211	800	2708

The reported smoking events are then used as input to the control theory model of smoking, in order to calculate the nicotine levels and threshold value during the 13 days. One 24 hour period was dropped because it was made of two half-days (one at the start and the other at the end of the data collection period). All the calculated data along with collected mobile data (eg. light, GPS Location, and activity labels etc.) are all combined to form the dataset tables for each participant. The labelled smoking events will be the labels for the data set. Fig. 3 shows the process of data collection.

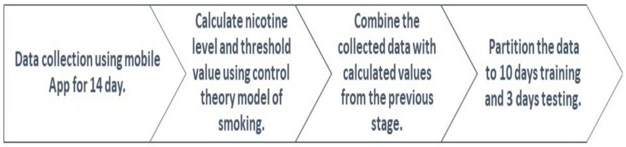
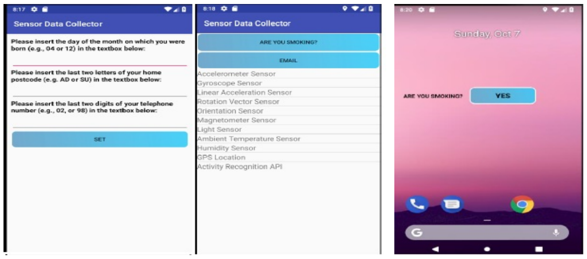

Data collection using mobile App for 14 day.
Calculate nicotine level and threshold value using control theory model of smoking.
Combine the collected data with calculated values from the previous stage.
Partition the data to 10 days training and 3 days testing.

Fig. 3: Overview of the study: data collection and processing steps.

A. Mobile App

Data collection took place using a mobile application developed for Android mobile users, using Android Studio (IDE). The main focus of the User Interface (UI) was to develop a user-friendly interface that provides no feedback to users, as so to avoid influencing their behaviour[23]. The UI was used to label smoking events, relying on participants' self-reporting of events. Users could report smoking events either by pressing a button on the main layout of the App, or by pressing a Widget on the home screen of the smartphone as can be seen in Fig. 4.



Please insert the day of the month on which you were born (e.g., 04 or 12) in the textbox below:

Please insert the last two letters of your home postcode (e.g. AD or SU) in the textbox below:

Please insert the last two digits of your telephone number (e.g., 02, or 98) in the textbox below:

(a) Participant registration screen (b) Application main layout (c) Amart phone home screeen

Fig. 4: Mobile application UI

The application was designed to run as a background service, which records data from the phone's sensors. This service was designed to restart itself whenever terminated (either by the OS or otherwise). This was implemented in order to overcome a new restriction forced by Android on the development of background services that run for long periods. The background service recorded one sample per minute from the sensed data. Collected data, along with smoking events were stored on an internal SQLite database.

B. Data collection

For this study, the participants were smoking adult over 18 years old, with a good level of English literacy. They each owned an Android mobile phone. Smokers are defined as those smoking at least 5 cigarettes a day. During the data collection period, the application was installed on the participant's smartphone for two weeks. No restrictions have been placed on their daily activities, and they have only been asked to report their smoking events and keep the GPS on. At this stage of the research data has been collected from 4 participants (3 females:1 male)[1].

Data were collected from several sensors in order to identify correlations between smoking events and the sensors reading. Table III shows the types of collected data. The goal is to use the collected data to find the association between smoking events and environmental data, in order to inform the implementation of a machine learning model that can automatically predict smoking events based on the occurrence of internal and external predictors. Following data collection, it emerged that not all sensors are available in all mobile models. Therefore the plan was modified to use only the common sensors that appear in most of the mobiles, i.e., the accelerometer and light sensors along with GPS values and human activity labels.

TABLE III: The number of labels in each of the three labelling categories.

Collected data group name	Description
ID	This is unique ID that Identify the user data, it is set by the user at the start of the study.
Timing value	This is time stamp DD-MM-YYYY,HH:MM:SS
Motion sensors data	Accelerometer, Gyroscope, Linear acceleration, Orientation, Rotation vector.
Environmental data	Magnetic field, Light level, Ambient temperature, Relative humidity, GPS location.
Activity labels	Google activity recognition API (Still, Running, Walking, Cycling, Tilting, and Driving).
Smoking labels	This is labelled by the user.

IV. Approach to model development

To design a machine learning model for smoking behaviour the control theory model of smoking will be combined with the decision tree classifier. At the start, each part of the model will be analysed separately before reaching the final model.

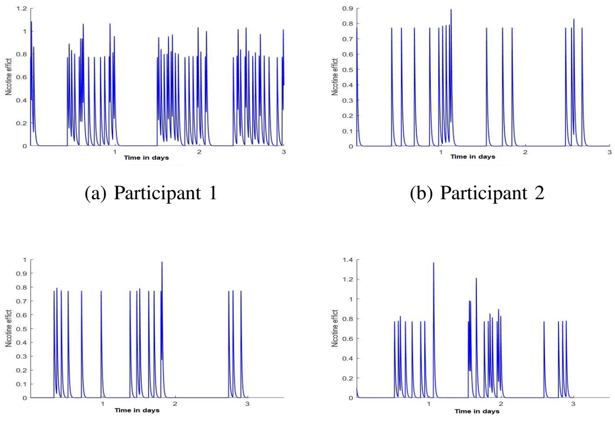
A. Control theory model of smoking

Using the reported smoking events, nicotine concentration was calculated using the control theory model of smoking [7] as shown in Fig. 5. Each peak in the figure represents smoking events, followed by a gradual decrease in the nicotine level until the next smoking event.

Fig. 6 shows the threshold values calculated using the control theory model. The peaks represent no smoking periods, the value of the threshold decreases by the increased number of cigarettes per day.

The control theory model also models the withdrawal and craving symptoms, Fig. 7 shows the values of withdrawal and craving over 10 days period.

[1] Although the number of participants appears small, a large volume of data was collected from each participant (approximately 1010 smoking events and 18720 samples each), making it sufficient for modeling a machine learning problem.



(a) Participant 1

(b) Participant 2

(c) Participant 3

(d) Participant 4

Fig. 5: Examples of 3 days of smoking behaviour by four participants, as modelled using control theory to represent nicotine levels.

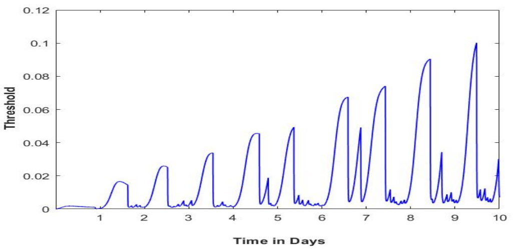
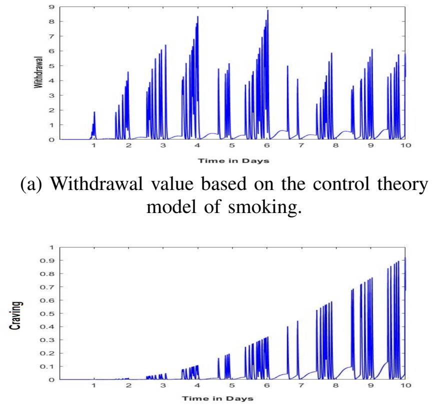


Fig. 6: Example of 10 days calculated threshold value using the control theory model of smoking and collected data from one of the participants.



(a) Withdrawal value based on the control theory model of smoking.

(b) Craving value based on the control theory model of smoking

Fig. 7: Example of 10 days calculated withdrawal and craving values using the control theory model of smoking and collected data from one of the participants.

B. *Classification of smoker behavioural data*

Three types of events occurr in the collected dataset, which are labelled as smoking (1), not smoking (2) and App app-off (0) events, where the later occur due to gaps in the dataset (e.g. participant turns the mobile off). Three types of Decision Tree models were explored; CART, Boosted Tree, and Tree Bagging (this selects a random subset of samples as in the random forest algorithm). The three classifiers are implemented and tested using the Matlab2017 "Statistics and Machine Learning Toolbox".

Initially, the classification methods are tested to see whether the classifier can detect the smokers events using only endogenous factors. Time, nicotine level, and threshold are used as input to the decision tree classifier. Table IV shows the classification accuracy, The data was tested using the iterative bootstrap process where three users are held for training and validation, and one participant is used for testing. The routine is repeated for each participant.

TABLE IV: The precision level of classification test based on only endogenous factors.

Calculated accuracy category	**The percentage accuracy level**		
	Tree Bagging	***Boosted Tree***	***CADT***
Participant 1 App off	4.7	2.6	1
Participant 1 smoking	70.2	83	65.6
Participant 1 not smoking	94.9	94.2	88.5
Participant 1 overall	54.46	20.8	27.59
Participant 2 App off	31.8	37.3	37
Participant 2 smoking	88	60.9	18.6
Participant 2 not smoking	64	66.9	66
Participant 2 overall	64	55.6	49.73
Participant 3 App off	39.3	17	17.4
Participant 3 smoking	73.1	77	43.7
Participant 3 not smoking	89.2	72.5	72.6
Participant 3 overall	68.28	43.08	42.73
Participant 4 App off	94.3	95.4	91.6
Participant 4 smoking	85.3	76.6	42
Participant 4 not smoking	15.6	16.8	17.7
Participant 4 overall	23.93	29.6	34.59
Average App off	42.525	38.075	36.75
Average smoking	79.15	74.375	42.475
Average not smoking	65.925	62.6	61.2
Average overall	52.6675	37.27	38.66

Secondly, to test the effect of adding the external factors on the performance of the classifier, GPS Location, light level, and human motion label are all used as predictors by the three classification methods along with the endogenous factors. Since the exogenous factors are personalized for each participant, the training model needs to be trained for each participant. The collected dataset for each participant was portioned into 10 days training (70% training and 30% validation) and 3 days testing. Table V shows the result of the testing process.

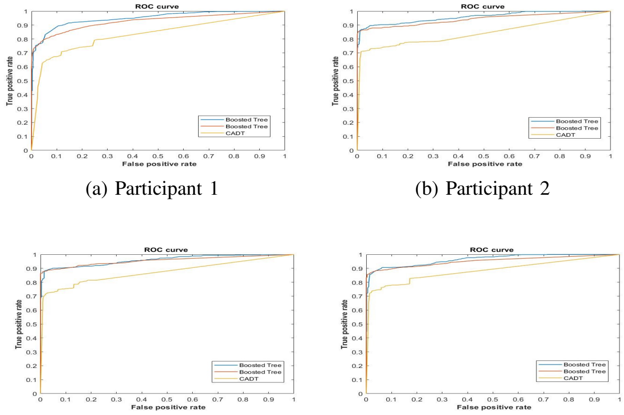
It can see from the tables that in general the performance of the Tree Bagging method is better than the other two classifiers, and that using all 6 predictors can give better overall performance. This can result in the conclusion that in order to model the smoker's behaviour the model has to

TABLE V: The precision level of classification test based all 6 predictors.

Calculated accuracy Category	The percentage accuracy level		
	CADT	***Boosted Tree***	***Tree Bagging***
Participant 1 overall	92.1	68	95.2
Participant 1 smoking	64.4	23.4	87.1
Participant 1 not smoking	97.0	97.7	96.1
Participant 1 unknown	0.0	16.8	87.5
Participant 2 overall	77.1	51.8	73.8
Participant 2 smoking	37.7	3.6	19.1
Participant 2 not smoking	85.5	89.8	95.5
Participant 2 unknown	69.7	64.9	68.1
Participant 3 overall	98.4	75.2	98.8
Participant 3 smoking	68	7.6	97.5
Participant 3 not smoking	99.5	99.8	98.1
Participant 3 unknown	100	97.1	100
Participant 4 overall	82.6	90.6	96.5
Participant 4 smoking	26.2	19.8	79.8
Participant 4 not smoking	19.8	11	61.3
Participant 4 unknown	99.3	96.8	98.8
Average overall	87.55	71.4	91.075
Average smoking	49.075	13.6	70.875
Average not smoking	75.45	74.575	87.75
Average unknown	67.25	68.9	88.6

be trained based on the individual behaviour for each person, and a general model will not target the unique needs that each person may have.

The ROC (Receiver Operating Characteristic) curve clarifies the differences in the performance between the three classifiers and shows how the performance increases when all the predictors are used. Fig. 8 and 9 compare the performance of the classifiers based on the classification methods and the number of input features, where the first figure shows the ROC curve for four participants using only the endogenous factors, while the second figure shows the classification performance for the four participants after considering all 6 predictors.



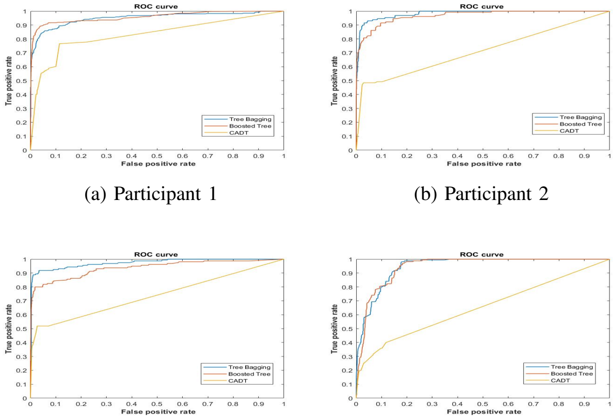
(a) Participant 1

(b) Participant 2

(c) Participant 3

(d) Participant 4

Fig. 8: Standerd ROC curves for smoking labels classification using only endogenous factors.



(a) Participant 1

(b) Participant 2

(c) Participant 3

(d) Participant 4

Fig. 9: Standerd ROC curves for smoking labels classification using 6 factors.

Other performance measures are displayed in TableVI;it can see from the table that the performance of the Bagging Tree is higher than the other classifiers.

TABLE VI: Performance indices for three classification methods.

Performance index	***Tree Bagging***	***Boosted Tree***	***CADT***
Precision	0.8858	0.7979	0.753735325
Recall	0.8117	0.77087	0.71498
F1 score	0.8282	0.7362	0.7129
Accuracy	0.9142	0.8678	0.8910

The bagging decision tree's ability to minimise the effect of the overfitting problem increased its performance over the other classification methods. Fig.10 shows the out-of-bag error against the number of classification trees grown.

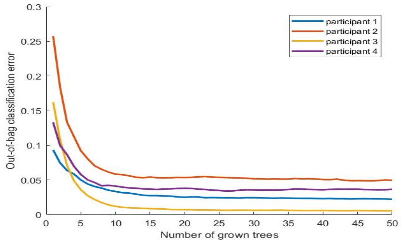


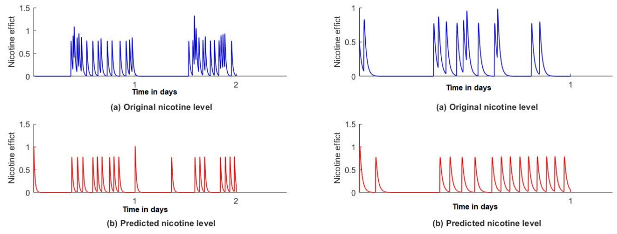
Fig. 10: Out-of-bag error against the number of classification trees grown.

V. Results

After testing the three classification methods, the Bagging Tree method was selected as a classifier to predict smoking events. The classifier predicts either smoking or non-smoking states, with the App off event being treated as non-smoking events. The point of the prediction to see if it can forecast the nicotine level (other than the original calculated values) using combined control theory and machine learning model.

The machine learning model combined with control theory model of smoking to model the smoker's daily behaviour in order to detect the smoking events using endogenous factors, and the other collected data (GPS Location, light level and human motion label). Since the exogenous factors are personalised for each participant, the training model needs to be trained for each participant. The data was tested iteratively, each participant data have been separated for twelve-day training and one-day testing, and then the routine is repeated for each day. This process helped in comparing the prediction level based on different day of the week.

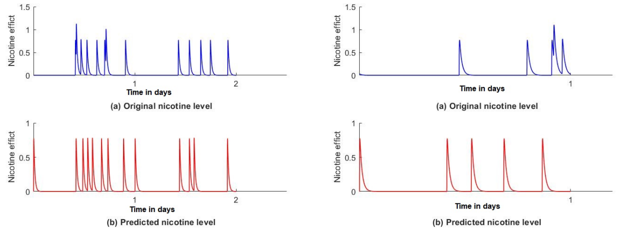
Fig. 11 and 12 shows the prediction result for two participants for randomly selected two regular weekdays along with the prediction of one of the weekend days for the same participant. All 6 predictors were used as input to the system. The nicotine level was predicted during the closed-loop process; no pre-calculated data was used.



(a) two randomly selected weekdays

(b) one weekend

Fig. 11: Example of predicted nicotine level for participant 1.



(a) two randomly selected weekdays

(b) one weekend

Fig. 12: Example of predicted nicotine level for participant 2.

Although some smoking events were missed, the model in general reliably models the smoking behaviour of each of the participants. The model is strongly relaying on the cooperation from the participants when reporting the smoking events accurately. The final design of model of the daily smoker's behaviour can be seen in Fig 13.

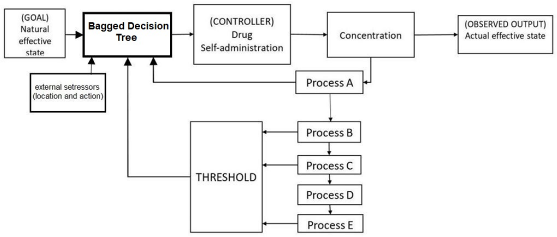


Fig. 13: Smoking behaviour model utilizing machine learning. Data are collected and processed using the steps described in Fig3. The 6 predictors are used as input to the Bagged decision tree classifier. A classification value of 1 represents a potential smoking event. This value is passed to the CONTROLLER, simulating the taking of a cigarette, and re-initializing the parameters of the control model to zero.

The results of the Mean Square Error (MSE), Root Mean Square Error (RMSE), and Normalized Root Mean Square Error (NRMSE), which are the error criteria used to measure the performance of the model, are displayed Table VII and VIII.

TABLE VII: The overall error rate of the proposed model over the regular days.

	MSE	***RMSE***	***NRMSE***
Participant 1	0.082667	0.287519	0.199763
Participant 2	0.038543	0.196323	0.124366
Participant 3	0.045966	0.214396	0.169778
Participant 4	0.027908	0.167057	0.121877
Average	0.048771	0.216324	0.153946

TABLE VIII: The overall error rate of the proposed model over the weekends.

	MSE	***RMSE***	***NRMSE***
Participant 1	0.084701	0.291034	0.202995
Participant 2	0.034426	0.185542	0.167896
Participant 3	0.039972	0.199929	0.203229
Participant 4	0.036116	0.190042	0.209595
Average	0.048804	0.216637	0.195929

VI. Conclusions

In conclusion, machine learning was sucessfully applied to model smokers' behaviour. The design model at this stage combines Bagged Decision Tree with the control theory model of smoking, and the results are generally promising. Six predictors of smokers' behaviour (nicotine effect level, the threshold value as calculated by control theory, light sensor, GPS location and type of activity) have been used to predict the smoking events. This design was able to adapt to the behaviour of individual smokers, but the accuracy of the smoking event prediction can still be improved.

It is expected that the accuracy of the system in predicting the smoking events will be increased by taking advantage of the information such as the indoor smoking ban in the UK and replacing the Google activity recognition by more accurate human behaviour classifier using the collected accelerometer values. It may also be possible to construct a combined model of individuals' behaviour, using additional external data such as the addresses of their work and home, and also public information on the location of businesses such as bars and resturants likely to be associated with smoking. These additons to the model are currently under consideration.

Acknowledgment

We thank the Manchester Metropolitan Unviersity intern students, employed under the funding program from the faculty of Health, Psychology and Social Care UKRIS and Faculty of Science and Engineering, who collected the data necessary for this study, and also the participants who gave up their time and battery power for this research.

References

[1] NHS, "Adult smoking habits in the uk: 2016," *office of national statistics*, June 2018.

[2] M. S. Fibla, U. Bernardet, and P. F. Verschure, "Allostatic control for robot behaviour regulation: An extension to path planning," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 1935–1942.

[3] J. M. Hughes, *Real World Instrumentation with Python: Automated Data Acquisition and Control Systems*. " O'Reilly Media, Inc.", 2010.

[4] B. Gutkin and S. H. Ahmed, *Computational Neuroscience of Drug Addiction*. Springer Science & Business Media, 2011, vol. 10.

[5] S. Metin and N. S. Sengor, "From occasional choices to inevitable musts: A computational model of nicotine addiction," *Computational intelligence and neuroscience*, vol. 2012, p. 18, 2012.

[6] R. L. Solomon, "The opponent-process theory of acquired motivation: the costs of pleasure and the benefits of pain." *American psychologist*, vol. 35, no. 8, p. 691, 1980.

[7] G. Bobashev, J. Holloway, E. Solano, and B. Gutkin, "Control theory model of smoking." *American psychologist*, 2017.

[8] K. P. Timms, D. E. Rivera, L. M. Collins, and M. E. Piper, "Control systems engineering for understanding and optimizing smoking cessation interventions," in *2013 American Control Conference*. IEEE, 2013, pp. 1964–1969.

[9] M. S. Businelle, P. Ma, D. E. Kendzor, S. G. Frank, D. W. Wetter, and D. J. Vidrine, "Using intensive longitudinal data collected via mobile phone to detect imminent lapse in smokers undergoing a scheduled quit attempt," *Journal of medical Internet research*, vol. 18, no. 10, 2016.

[10] R. S. Schick, T. W. Kelsey, J. Marston, K. Samson, and G. W. Humphris, "Mapmysmoke: feasibility of a new quit cigarette smoking mobile phone application using integrated geo-positioning technology, and motivational messaging within a primary care setting," *Pilot and feasibility studies*, vol. 4, no. 1, p. 19, 2018.

[11] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2387–2403, 2013.

[12] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[13] N. D. Vanli, M. O. Sayin, M. Mohaghegh, H. Ozkan, and S. S. Kozat, "Nonlinear regression via incremental decision trees," *Pattern Recognition*, vol. 86, pp. 1–13, 2019.

[14] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[15] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[17] M. C. Tu, D. Shin, and D. Shin, "A comparative study of medical data classification methods based on decision tree and bagging algorithms," in *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 2009, pp. 183–187.

[18] W. Zhu, M. Xie, and J.-F. Xie, "A decision tree algorithm for license plate recognition based on bagging," in *Wavelet Active Media Technology and Information Processing (ICWAMTIP), 2012 International Conference on*. IEEE, 2012, pp. 136–139.

[19] P. Bonissone, J. M. Cadenas, M. C. Garrido, and R. A. Díaz-Valladares, "A fuzzy random forest," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 729–747, 2010.

[20] G. J. Briem, J. A. Benediktsson, and J. R. Sveinsson, "Multiple classifiers applied to multisource remote sensing data," *IEEE transactions on geoscience and remote sensing*, vol. 40, no. 10, pp. 2291–2299, 2002.

[21] M. A. Mim and K. S. Zamil, "Gis-based analysis of changing surface water in rajshahi city corporation area using support vector machine (svm), decision tree & random forest technique," *Machine Learning Research*, vol. 3, no. 2, p. 11, 2018.

[22] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[23] S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M. P. Eccles, J. Cane, and C. E. Wood, "The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions," *Annals of behavioral medicine*, vol. 46, no. 1, pp. 81–95, 2013.