Analyzing Smoking Behavior: A Comparative Study of Neural Networks and Random Forests

Herbert Shin
Faculty of Engineering
University of Western Ontario
London, Canada
hshin63@uwo.ca

Eyoel Hailemariam
Faculty of Science
University of Western Ontario
London, Canada
ehailema@uwo.ca

Thai Luong
Faculty of Science
University of Western Ontario
London, Canada
tluong26@uwo.ca

Colin Smith
Faculty of Engineering
University of Western Ontario
London, Canada
csmit586@uwo.ca

***Abstract*— The impact of smoking on human health is a subject of extensive research. However, the early detection and classification of individuals based on their smoking habits remain a challenge. This project aims to address this issue by implementing machine learning algorithms and neural networks to analyze body signals and classify individuals as smokers, prior smokers that quit, or non-smokers. The plan involves collecting a diverse dataset sourced from Kaggle which contains various physical, biological and psychological features associated with smoking habits. The AI model will be trained using a deep feedforward neural network and random forests model. Furthermore, feature engineering and data augmentation techniques will be employed to enhance the model's performance and robustness. The ultimate goal is to develop a precise AI system for detecting smoking behavior, which could greatly impact public health and personalized healthcare. This project contributes to AI in healthcare and paves the way for future studies to enhance model sophistication and signal diversity for better accuracy. Integrating AI into public health could lead to more tailored and proactive healthcare solutions, reducing the impact of smoking-related diseases.**

Index Terms—Neural Network, Random Forest, Representational Dissimilarity Matrices, Multidimensional Scaling, Receiver Operating Characteristic, Rectified Linear Unit, Area Under the ROC Curve.

I. Introduction

The detrimental effects of smoking on human health are well-documented, with smoking being a leading cause of preventable morbidity and mortality worldwide [1]. While the impact of smoking on health is extensively researched, early detection and classification of individuals based on their smoking habits pose significant challenges [2]. Accurate identification of smokers can aid in targeted interventions and personalized healthcare, potentially reducing the burden of smoking-related diseases.

To address this challenge, this project explores both a neural network and random forests solution. These models are trained on a comprehensive set of features to precisely classify individuals as smokers, non-smokers, or former smokers. The selection of these two distinct approaches allows for a comparative evaluation, highlighting the neural network's ability to capture complex patterns and its susceptibility to overfitting, against the random forest model's strength in generalization and robustness.

The integration of this AI system into public health strategies could revolutionize smoking cessation interventions, leading to more effective treatment programs and a reduction in smoking-related diseases, ultimately contributing to the global efforts in reducing the burden of smoking-related diseases.

II. Related work

Recent studies have explored machine learning approaches to classify individuals based on smoking status. Frank et al. [3] employed machine learning algorithms to predict smoking status based on blood test results, with the Logistic algorithm showing high accuracy. Davagdorj et al. [4] used machine learning techniques to predict success in smoking cessation interventions, highlighting the importance of factors such as age and alcohol use.

Despite these advancements, there is a research gap in the comprehensive evaluation and comparison of different AI models. Specifically, through the use of neural networks and random forests to analyze smoking-related behaviors. Existing studies often focus on specific biomarkers or psychological factors, and there is a need for a more holistic approach that evaluates the performance between machine learning models and neural networks.

This project aims to address this gap by using both neural networks and random forests to analyze a diverse dataset containing physical, biological, and psychological features associated with smoking habits. This comparative analysis will provide insights into the impacts of each solution and contribute to the development of more accurate and reliable AI systems for detecting smoking behavior, ultimately impacting public health and personalized healthcare.

III.Data

For solving the problem of identifying smokers based on their biological, physical, and psychological features, a dataset is required for training and testing the models. The dataset freely available from Kaggle is “Smoking and Drinking Dataset with body signal” [5]. The dataset is collected from the National Health Insurance Service in Korea and all personal information and sensitive data were excluded. The dataset features 991346 entries with each entry having parameters shown in Table 1. This data set is broken down with 85% being used for training, 7.5% used for validation, and 7.5% used for testing.

TABLE I
Parameters of Dataset

Parameter Name	Parameter Description
Sex	Male, female
age	Round up to 5 years
height	Round up to 5 cm
weight	kg
sight_left	Eyesight (left)
sight_right	Eyesight (right)
hear_left	Hearing left, 1 (normal), 2 (abnormal)
hear_right	Hearing right, 1 (normal), 2 (abnormal)
SBP	Systolic blood pressure [mmHg]
DBP	Diastolic blood pressure [mmHg]
BLDS	BLDS or FSG (fasting blood glucose) [mg/dl]
tot_chloe	Total cholesterol [mg/dL]
HDL_chole	HDL cholesterol [mg/dL]
LDL_chole	LDL cholesterol [mg/dL]
triglyceride	Triglyceride [mg/dL]
hemoglobin	Hemoglobin [g/dL]
urine_protein	Protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
serum_creatinine	Serum (blood) creatine [mg/dL]
SGOT_AST	SGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L]
SGOT_ALT	ALT(Alanine transaminase)[IU/L]
gamma_GTP	y-glutamyl transpeptidase[IU/L]
SMK_stat_type_cd	Smoking state, 0(never), 1(used to smoke but quit), 2(still smoke)
DRK_YN	Drinker or Not

IV. Data preprocessing

Before the data can be used for developing models there are a few simple operations to be done to the dataset.

A. *Removing DRK_YN*

The first step is to remove the last column of the dataset that is related to if a person is a drinker or not. This is because the original data set was intended for determining both if a person smoked and drank but the purpose of this project only focuses on smoking.

B. *Removing noise and splitting dataset*

All noisy data points that have missing information in them are dropped from the dataset. After that, the data is split into input and output and changed to a numpy array.

C. *Transforming data*

The function get_dummies() is used from pandas library to convert the ‘sex’ column into binary where 0 is female and 1 is male.

Then, all the remaining data is squished using StandardScaler(). StandardScaler() is a transforming function that calculates the mean and standard deviation of each feature in the dataset and then transforms the features by subtracting them with their mean and dividing them by the standard deviation. After that, the output data is transformed and stored as an integer using the LabelEncoder() function.

D. *Separating data*

Finally, the data is separated into three categories, training data (85%), validation data(7.5%), and testing data(7.5%), where each category has approximately the same distribution of output as the original dataset, which can be seen in figure 1. This data can now be inputted into the model for the corresponding step of developing the model.

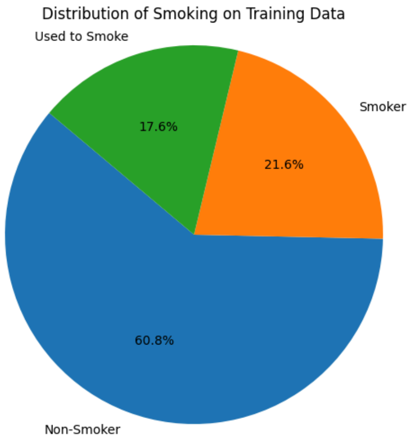


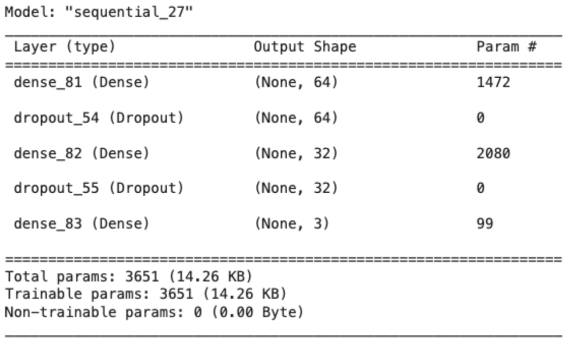
Fig. 1 Distribution of Smoking on Training Data

V. Methods

The project aims to create a neural network model that can detect smoking behavior based on physical, biological, and psychological features. This model will be evaluated for its accuracy in predicting smoking behaviors as well as its training and operational requirements. A machine learning random forest model will also be created and evaluated for the same parameters. These two models will be compared to each other to assess the benefits of each model.

A. NN Model

A Neural Network model is the most widely recognized and used machine learning model that is loosely inspired by the structure and function of a human's brain. An NN model is also very useful for our project since it Our custom NN model consists of Dense layers with ReLU activation function and Dropout layers. The summary of our custom NN model looks like this:



```
Model: "sequential_27"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_81 (Dense)            (None, 64)                1472

 dropout_54 (Dropout)        (None, 64)                0

 dense_82 (Dense)            (None, 32)                2080

 dropout_55 (Dropout)        (None, 32)                0

 dense_83 (Dense)            (None, 3)                 99

=================================================================
Total params: 3651 (14.26 KB)
Trainable params: 3651 (14.26 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 2 Summary of our NN Model

We chose Dense layers (fully connected layers) for our model since each neuron in one layer is connected to every neuron in the subsequent layer, the model will be able to learn complex patterns and relationships in the data.

We chose to have Dropout layers after each hidden layer as a regularization technique so that we can prevent the model from overfitting to the training data. We chose a dropout rate of 0.2 which means that 20% of the input units will be randomly set to zero during each training iteration. By doing this, the model will learn more robust and generalizable features from the data since it must learn to perform well even when some of the units are missing. We chose a dropout rate of 20% since our dataset is simple and does not need as much regularization to generalize unseen data well.

Since our output is represented as multi-class classification, for the last output layer, we gave it 3 nodes and made the activation function softmax. We chose softmax activation function for the last layer because it ensures that the sum of the output probabilities across all classes is equal to 1 which makes it interpretable as probabilities. This is crucial because it allows us to interpret the output as the probability that a given input belongs to each class.

The hyper parameters we chose for our NN model are:

- Number of Layers: 3
- Number of Neurons per Layer: 64, 32, 3
- Activation Functions: ReLU (for the first two), SoftMax (for last)
- Dropout Rate: 20%
- Learning Rate: Dynamic. Since for the model's optimizer we are using 'adam', at first the Learning Rate is set to 0.001, but during training, it adapts dynamically based on the statistics of gradients for each parameter.
- Batch Size: 12,800
- Optimizer: Adam
- Loss Function: Sparse Categorical CrossEntropy. This loss function computes the cross entropy loss between the predicted probability distribution that is obtained from the softmax activation function in the output layer and the truth class label. We used this loss function because it is suitable for our type of classification task.
- Initialization Method: Uses TensorFlow's default initializer, which is Glorot uniform initialization. This method initializes the weights with values drawn from a uniform distribution that is bounded by a range which is determined by the number of input and output units.
- Regularization Techniques: Dropout

During one Epoch, the neural network iterates over the entire dataset in batches, computes the loss for each batch, calculates the gradients with respect to the loss, and then updates the weights and biases using Gradient Descent. We experimented with different amounts of Epoch and recorded the training and validation accuracy and loss on each Epoch. The results can be seen in Fig 3 and Fig 4.

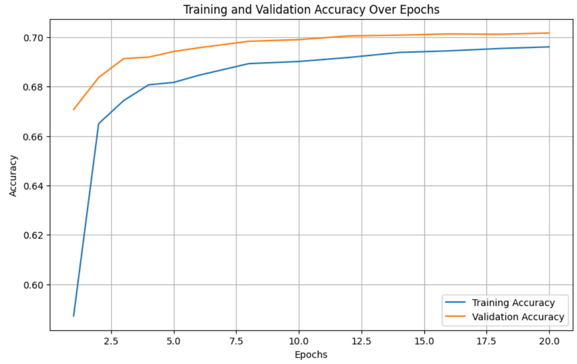


Fig. 3 Training and Validation Accuracy Over Epochs

- The accuracy on both training and validation growth logarithmically. Around Epoch of 3, the accuracy starts to stabilize around 70%

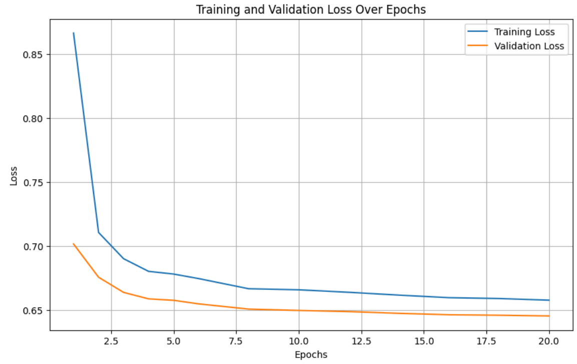


Fig. 4 Training and Validation Loss Over Epochs

- The loss on both training and validation decreases logarithmically. Around Epoch of 3, the loss starts to stabilize at around 65%

B. *Random Forest Model*

The random forest model is a common machine learning model that can produce effective results without requiring extensive tuning. The model operates by making a collection of decision trees based on the different input features, and then averages the resulting decisions across all the trees to produce the final decision. The model is capable of assigning levels of importance to different features based on how impactful that feature is to changing the final decision. Additionally, the random forest model avoids overfitting as long as a sufficient number of trees are used in the model. The issues of this model come from the more trees used, while making more accurate predictions, take longer to make the predictions. These models are fast to train, but slow to use as the number of trees increases [6].

This model will be evaluated primarily on its accuracy, but other factors such as speed of use when given data to make a decision will also be evaluated to assess the full value of the model when compared to the NN model.

VI. Results

The results of this project reveal insightful comparisons between the performance of Neural Networks (NN) and Random Forest (RF) models in classifying individuals based on individual smoking status from an open source database. Utilizing a dataset with diverse physical, biological, and psychological features, both models were meticulously trained and evaluated on their predictive capabilities.

A. *NN Model*

The confusion matrices for Validation set and test set result in a few patterns emerging in the models' predictive performance. Validations set showed a strong capability to correctly identify class 0 instances as confirmed by the prominent positive counts in the confusion matrix. However, there is a noticeable decrease between class 2 (current smoke) and class 1 (used to smoke) , suggesting that these classes may share similar characteristics or that the model hasn't learned to distinguish them effectively from class 0 (non-smoker). The same trend occurs in the test set, with some additional errors in identifying class 0, hinting that the model may have overfitted to the training data or that a more varied dataset might be required to improve its ability to generalize to new data.

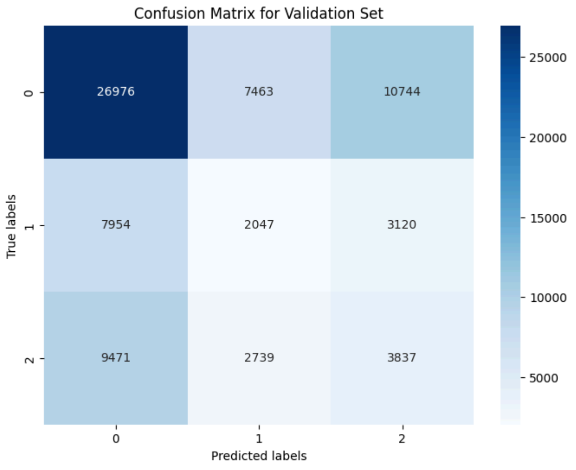


Fig. 5 Confusion Matrix for NN on validation set.

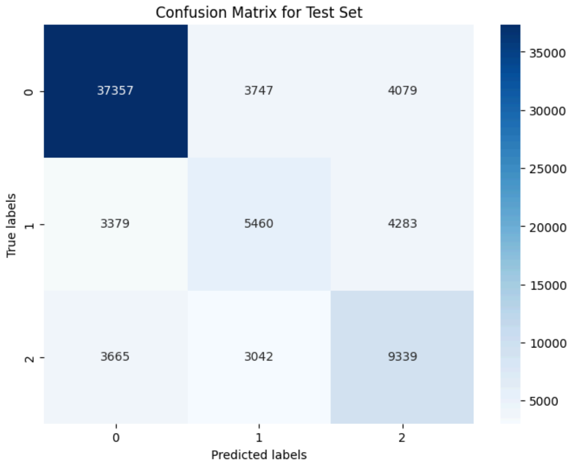


Fig. 6 Confusion Matrix for NN on test set

The RDM and MDS were created using 200 testing samples.

The RDM plot displays a matrix where each element in the matrix is the Euclidean distance between the activation vectors of the corresponding inputs.

The MDS plots, which visualize the data in a simplified 2D space, result in some separation between classes, even if it is not distinct. It shows that the activations for class 0 inputs are more distinct than classes 1 and 2, where there is not a clear division, aligning with the overlap that we have seen in the classification results

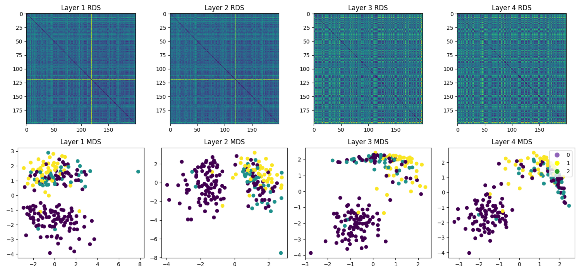


Fig. 7 RDM and MDS Visualization of the Activations of Each Layer

The table II presents a classification report for a test dataset, detailing the precision, recall, and F1-score for three different classes (0, 1, and 2).

-Class 0, the non-smokers, is predicted with high precision and recall, both at 0.84 and 0.83, respectively, reflecting a strong model performance for this group.

-Class 1, representing the used to smoke group, shows considerably lower precision and recall, at 0.45 and 0.42, indicating the model struggles to correctly identify this class.

-Class 2, the regular smokers, has moderate precision and recall values at 0.53 and 0.58.

The macro and weighted averages suggest that while the model is reasonably accurate overall (0.70 accuracy), its ability to predict across classes uniformly (macro averages) is less effective. This discrepancy highlights a need for model improvement or training adjustments, particularly for classes 1 and 2. The 'support' column indicates the number of true instances for each class in the test data, showing a class imbalance that may affect the model's learning capability.

TABLE II
TEST CLASSIFICATION REPORT

	Precision	**Recall**	**f1-score**	**support**
Class 0	0.84	0.83	0.83	45183
Class 1	0.45	0.42	0.43	13122
Class 2	0.53	0.58	0.55	16046
Accuracy	x	x	0.70	74351
Macro avg	0.60	0.61	0.61	74351
Weighted avg	0.70	0.70	0.70	74351

Using ROC- Receiver Operating Characteristic curve illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is used to assess the trade-offs between true positive rate (sensitivity) and false positive rate (1-specificity).

Its quantification through the AUC (Area Under the ROC Curve) provided a singular measure of performance, encompassing all possible classification thresholds. This metric was crucial in assessing

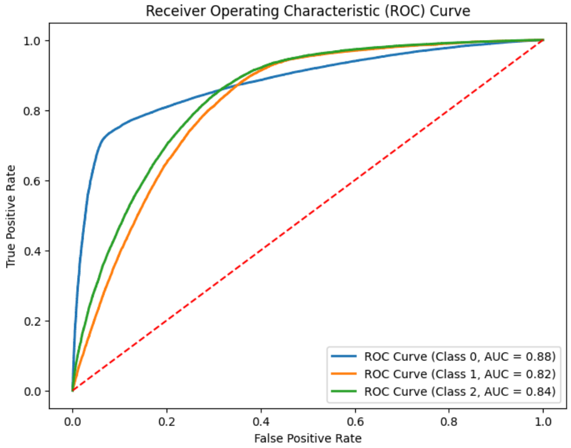


Fig. 8 NN AUC and ROC Curve on Test Data

The bar graph depicts a consistent pattern across the classes.

-Class 0 demonstrates a higher F1 Score and Test Accuracy, indicating a stronger performance of the model in accurately identifying instances of this class.

-Classes 1 and 2 exhibit similar levels of F1 Scores and Test Accuracies, which are notably lower than those of Class 0, suggesting that the model's ability to correctly classify these instances is less robust.

The disparity between the F1 Score and Test Accuracy within each class is minimal, underscoring the model's balanced precision and recall. This balance is essential in contexts where both false positives and false negatives carry significant consequences.

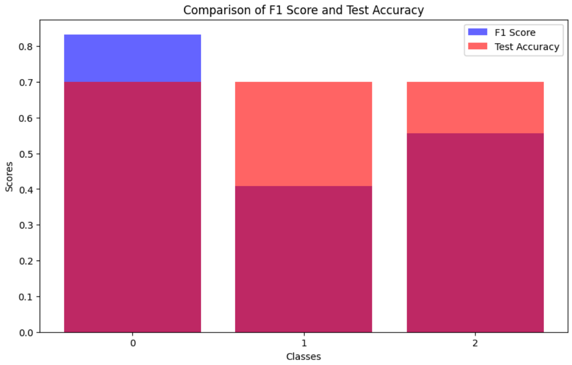


Fig. 9 Comparison of F1 Score and Test Accuracy of NN model

B. *Random Forest Model*

Both NN and RF models show some differences in their classification behaviors. The RF model shows a slightly better balance in classifying classes 1 and 2, indicating a more generalized model. The RF model seems to be less biased towards class 0 and is better at distinguishing between all three classes, albeit with some overlap between classes 1 and 2. However, this comes at a slight cost to the correct classification of class 0 compared to the NN model.

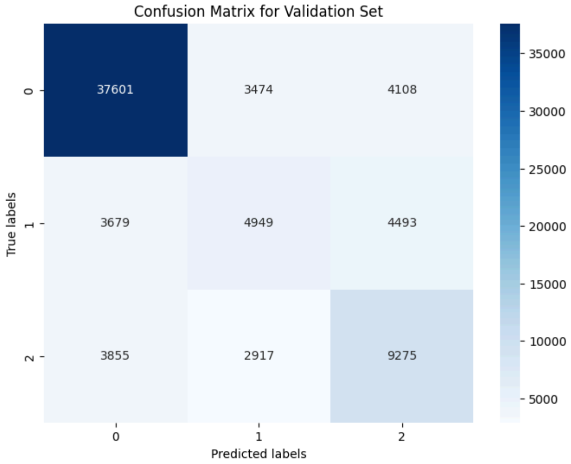


Fig. 10 Confusion Matrix for RF on validation set.

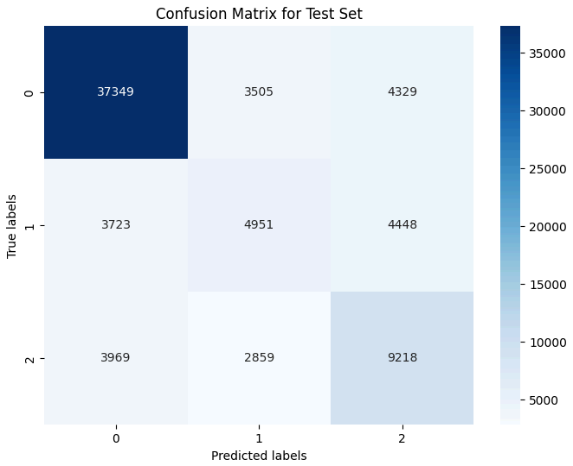


Fig. 11 Confusion Matrix for RF on test set.

Figure 8 of NN results for ROC curve show slightly higher values for class 0 and class 2 which could be attributed to its ability to model complex relations in the data through multiple layers and non linear activation.

However, the RF model is not so different with approximate AUC values that suggest it is also having a high true positive rate while keeping a low false positive rate.

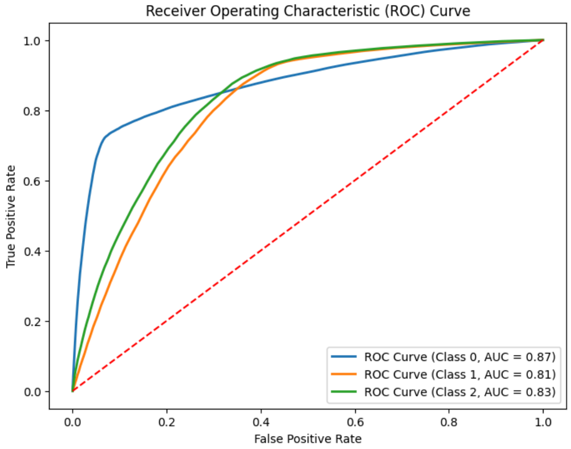


Fig. 12 RF AUC and ROC Curve on Test Data

VII. Discussion

The discernibility between class 1 and class 2 within the neural network's MDS plots underscores an intrinsic challenge in the classification process. These two classes 1 and 2, potentially representing varying degrees of smoking habits such as occasional versus habitual smokers, exhibit significant overlap. This could be attributed to the subtle physiological differences captured by the dataset or the similar patterns in feature space, confounding the network's ability to effectively distinguish between the two. Moreover, the disproportionate representation of class 0, the non-smokers, in the training set could bias the model towards more accurately identifying non-smokers at the expense of the nuanced differentiation required for the other classes.

A potential avenue for improving clarity between classes 1 and 2 might involve balancing the dataset to mitigate the dominance of class 0 or incorporating additional distinct features that could be more telling of the differences between occasional and habitual smokers. Additionally, advanced techniques such as ensemble learning methods, which combine multiple models, or exploring unsupervised learning to better understand underlying patterns, could offer new insights. Such strategies may foster a more nuanced recognition of smoking behaviors, enhancing the model's ability to generalize across diverse smoking profiles and leading to more personalized healthcare recommendations.

VIII. Conclusion

In conclusion, this project's exploration of a neural network and a random forest model for classifying individuals based on their smoking status revealed distinct strengths and limitations. The neural network model demonstrated strong performance in accurately identifying non-smokers but faced challenges in distinguishing between current and former smokers. On the other hand, the random forest model showed a more balanced classification across all three classes, suggesting its superior generalization capabilities. Future work will involve refining the models by exploring advanced techniques and additional features to improve their discrimination between smoking behaviors. This project highlights the importance of model selection based on class-specific performance and the potential of machine learning in public health applications, paving the way for future research to develop more accurate and reliable AI systems for smoking behavior detection.

References

[1] C. J. L. Murray et al., "The Global Burden of Disease Attributable to Tobacco Smoking," Journal of the American Medical Association, vol. 292, no. 10, pp. 1152-1157, 2004. (accessed Apr. 8, 2024).

[2] World Health Organization, "WHO Report on the Global Tobacco Epidemic," Geneva, Switzerland: World Health Organization, 2019. (accessed Apr. 8, 2024).

[3] C. Frank, A. Habach, and R. I. Seetan, "Predicting smoking status using machine learning algorithms and statistical analysis," Journal of Computing Sciences in Colleges, vol. 33, no. 4, pp. 120-127, Mar. 2018. (accessed Apr. 8, 2024).

[4] K. Davagdorj, J. S. Lee, K. H. Park, and K. H. Ryu, "A machine-learning approach for predicting success in smoking cessation intervention," International Journal of Environmental Research and Public Health, vol. 16, no. 19, p. 3654, 2019. (accessed Apr. 8, 2024).

[5] Soo.Y, “Smoking and drinking dataset with body signal,”Kaggle,https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/data?select=smoking_driking_dataset_Ver01.csv (accessed Apr. 8, 2024).

[6] “Random Forest: A complete guide for machine learning,” Built In, https://builtin.com/data-science/random-forest-algorithm#how (accessed Apr. 8, 2024