Smoking Status Prediction from Bio-Signals

1st Jie Min
Master of Data Analytics
University of Western Ontario
London, Canada
jmin67@uwo.ca

2nd Huilin Niu
Master of Biochemistry
University of Western Ontario
London, Canada
hniu32@uwo.ca

3rd Jonathan Oxman
Master of Computer Science
University of Western Ontario
London, Canada
joxman@uwo.ca

4th Muxuan Sun
Master of Data Analytics
University of Western Ontario
London, Canada
msun284@uwo.ca

5th Yifu Zheng
Master of Data Analytics
University of Western Ontario
London, Canada
yzhen487@uwo.ca

***Abstract*—This paper explores the application of machine learning techniques to predict smoking status based on bio-signals, providing a non-intrusive approach to identifying smoking behavior among unresponsive patients. Bio-signals, particularly as collected from blood and bodily fluid samples, were used as predictors. Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Neural Networks (NN) were employed to classify smoking status. We evaluated each model's performance in terms of accuracy, sensitivity, and specificity to determine the most effective approach. NN and LR offer fairly accurate predictions with comparable performance (AUC=80%), while RF performed significantly better (AUC=87%). Our findings suggest that machine learning models, particularly RF, hold promise for identifying smoking status from bio-signals, paving the way for real-time health monitoring applications.**

***Index Terms*—patient monitoring, regression analysis, support vector machines, neural networks, random forests**

I. Background

Tobacco use and smoking are the leading cause of preventable deaths in the USA, taking responsibility for over 20,000,000 preventable American deaths between the years 1965 and 2014 [1]. In particular, habitual smoking is a comorbidity that can cause complications in many medical procedures, particularly surgical operations. The main effects of smoking come from nicotine: these include various forms of carcinogenesis, increased glycogen synthesis, and gastric reflux [2]. However, smoking also exposes the body to other chemicals, such as nitrosamines, which can act as an immunosuppressant, and ammonia, which is corrosive and can cause respiratory tract inflamation. The presence of these substances in the bloodstream can cause unexpected interactions with anesthesia medicines used in surgery. [3]

Identification of smokers given a cursory medical examination can offer crucial insights into potential avenues of care. However, ascertainment of smoking status is not a ubiquitous practice among hospitals [4]. The seemingly simplest approach to determining smoking status is direct questioning of the patient: however, this option is not available if the patient is unconscious or otherwise unresponsive, which is often the case when emergency surgery is needed. Moreover, it is possible and indeed not unlikely for a responsive patient to attempt to mislead healthcare providers on this topic, with one study indicating that 17% of patients lie about smoking [5].

II. Objectives & Hypothesis

We aim to solve the smoker status prediction question by creating a solution for automatic smoking status prediction without relying on information reported by the patient. High-quality patient bio-signal data is readily available in large quantities, so a machine learning approach is well-suited to the problem. The hypothesis is that a machine learning approach will yield a model with strong predictive power that is able to identify likely smokers with minimal effort on the healthcare provider's part.

III. Data Pre-Processing

We consider a dataset consiting of 23 features from 38984 samples [6]. The target feature is the smoking status, while the remaining 22 features are explanatory. These are as follows:

- Basic physical characteristics: Age, height, weight, waist circumference, and presence of dental caries.
- Sensory capabilities: Test results for both sides of eyesight and hearing.
- Indicators for anemia and diabetes: Levels of hemoglobin and blood sugar.
- Indicators for heart disease: Levels of fat in the bloodstream, including triglyceride, HDL and LDL as well as a total measurement.
- Indicators for liver and kidney damage: Levels of protein in urine, creatinine in serum, and AST, ALT (types of glutamic oxaloacetic transaminase), and γ-GGT in blood.

The data was pre-cleaned with no missing values, with a 60%-20%-20% train, validation, and test split. The only modification we made to the dataset was to group ages by ranges representing discrete age groups. The bins for these are as follows: 20-32, 33-45, 46-58, 59-71, 72-85. As the quantity of data is high, we can afford to further split the test set equally for validation and testing. This large validation set helps us narrow the confidence interval for the generalization error of the model.

IV. Analysis & Visualization

As this is a classification problem, several methods naturally suggest themselves. We approach the dataset from several directions to compare these methods. The most basic approach is Logistic Regression (LR), which we employ for feature selection and to give a baseline for model performance. We also consider other choices for models, including Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN). To compare the approaches, we compute each ROC curve and the corresponding AUC. There is an additional nuance in this classification problem: the damage done by a false negative (FN) as opposed to a false positive (FP). In our application, an FP is unlikely to cause major problems, but as certain medical procedures are extremely dangerous for smokers, FNs should be penalized heavily for our models.

A. *Baseline Model: Feature Selection Via Logistic Regression*

We first consider the most direct approach to our classification problem: logistic regression. We begin by training the model on the full dataset with all the features. For feature selection, we determine the importance of each feature by computing the log odds (obtained from a transformation of the actual probabilities, for normalization) (Figure 1). For further visualization and analysis of multicolinearity, we also compute the correlation matrix for all the features (Figure 2). From these visualizations, we see that several features have weights of very low magnitude in the model: in particular, the features related to hearing and urine protein levels. Moreover, looking at the correlation matrix, we see that LDL and Cholesterol levels are only strongly correlated with each other, and have no correlation with the rest of the model. (The same holds true for the left and right hearing features.) Therefore, we conclude these features are unlikely to be beneficial in the classification problem and drop them.

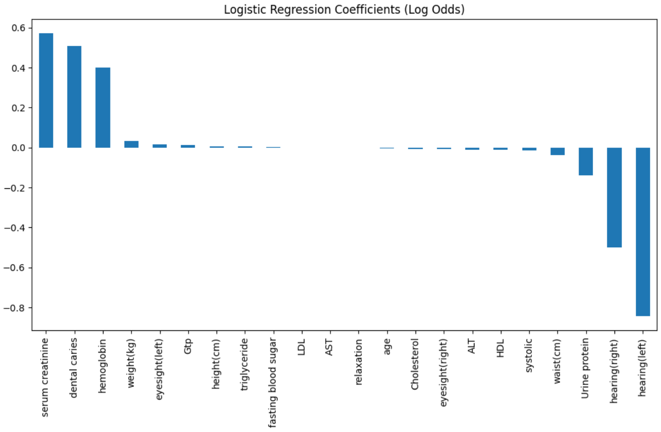

Fig. 1. Logistic Regression Feature Significance

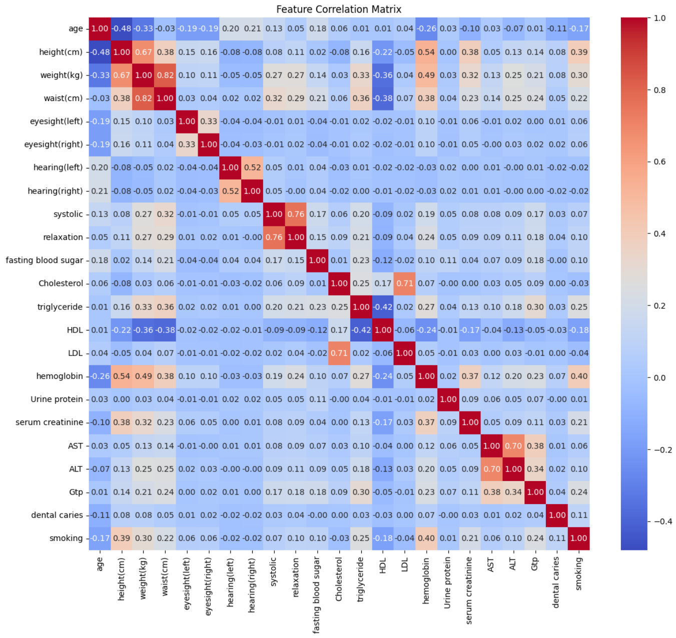

Fig. 2. Correlation of Features

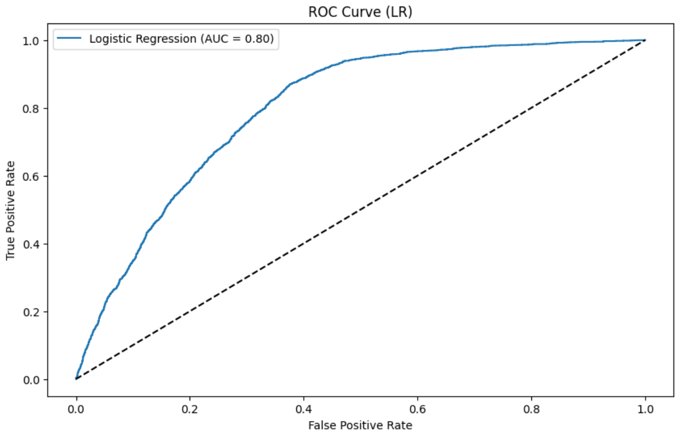

Fig. 3. ROC Curve for Logistic Regression

B. *Logistic Regression*

Having selected our features, we now conduct the basic logistic regression. First, we tune the hyperparameters on the validation set via grid search with cross-validation. We consider five candidates for regularization strength: $0.01, 0.1, 1, 10, 100$. Having already performed the feature selection, there is little benefit to applying an L_1 loss function, and so we only consider the L_2 case. After performing five-fold cross-validation, we conclude the best parameter for regularization strength is 1, with an cross-validation accuracy score of 0.81. Fitting the model and testing it on the test set, we obtain an accuracy of 0.72 and the ROC curve (Figure 3).

Note the high AUC score of 0.80, which indicates a good specificity and sensitivity. To analyze the specificity and sensitivity further, we consider the confusion matrix (Figure 4).

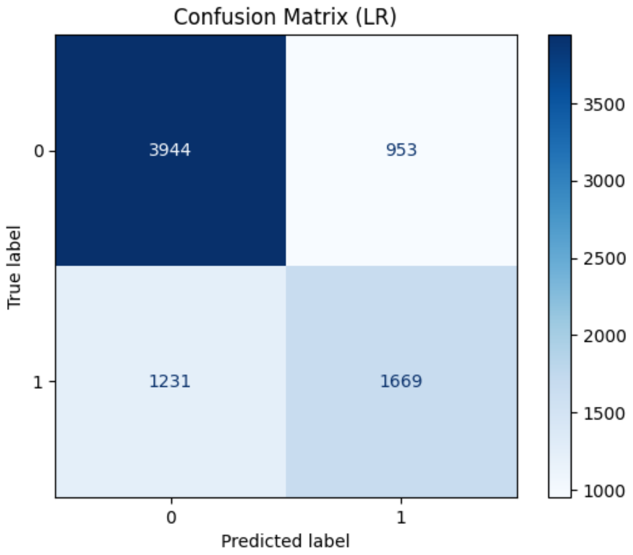

Fig. 4. Confusion Matrix for Logistic Regression

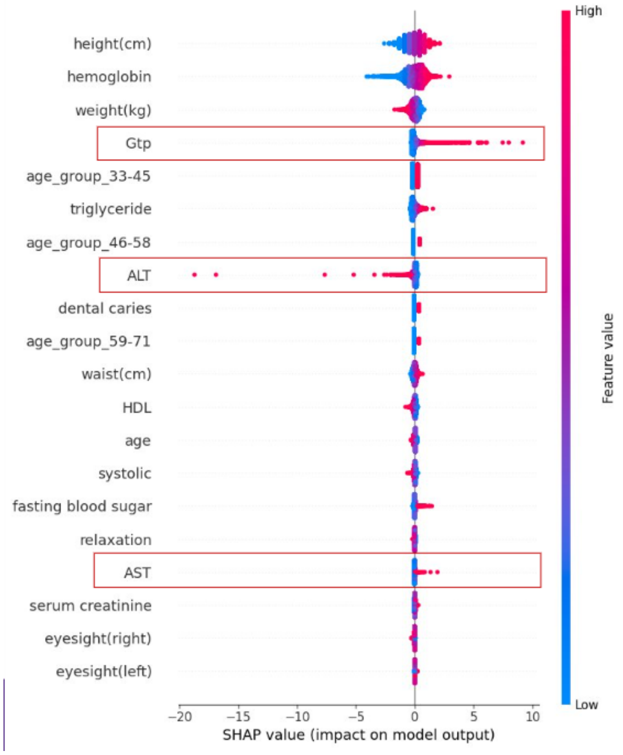

Fig. 5. SHAP Plot for Logistic Regression

Here, we see that false negatives on the test set are the least common outcome, with 1231 instances, which makes up about 0.16 of the cases. To visualize the importance of each bio-signal in the prediction, we consider the SHapley Additive exPlanations (SHAP) plot (Figure 5).

From this, it can be seen that the features with the most significance in the prediction of smoking status in our LR model are the height(cm), hemoglobin, weight(kg), and Gtp, ALT, and AST features giving it a significant push towards either predicting non-smoking/smoking.

C. Support Vector Machine

The second predictive model we consider is a linear SVM. Here, the optimal hyperparameters were determined by grid search with cross-validation to be a regularization constant of 10 (chosen from 0.1, 1, 10, and 100) and the penalty to be L_2 (as opposed to L_1). The extimated cross-validation accuracy is 0.81. After fitting the model and testing it on the test set, we obtain an accuracy of approximately 0.73 and the ROC curve (Figure 8). (For brevity, these figures will be shown in Appendix A.) The AUC score here, too, is 0.80. We again consider the confusion matrix (Figure 9). This yields 1223 false negatives, very comparable with the LR results and still occurring in the test set with a probability of about 0.16. We again consider the SHAP plot (Figure 10). From this, we can see that in this model as well, the most impactful features are height(cm), hemoglobin, and age_group_33-45, while similar Gtp, ALT, and AST giving the most significant push.

D. Random Forest

Next, we consider a random forest model based on binary estimators obtained from each of our (non-dropped) features. Once again, we tune the hyperparameters on the validation set and obtain (from several lists of options, omitted here for brevity) the optimal parameters to be 100 estimators, a maximum forest depth of 20, the minimum number samples to split on to be 2, the minimum number of samples to base a leaf on to be 1, and the square root function to be used when determining the number of subsampled features. Performing five-fold cross-validation, we achieve a cross-validation accuracy score of 0.830. Fitting the model and testing it on the test set, we obtain an accurancy of 0.79 and the ROC curve (Figure 11). Moreover, the AUC score is very high at 0.87. The confusion matrix (Figure 6) shows the accuracy. Compared to the basic logistic regression approach, we see here that the rate of both FPs and FNs is lower: in particular, the rate of FNs is 0.11 which is a clear improvement. Since this model has so many trees, it is not computationally feasible to produce a SHAP plot. Instead, we use a feature importance plot, which gives a less accurate but still reasonable summary of which features impact the model predictions the most significantly (Figure 7).

We can see that the three main predictors in the previous SHAP plots also show up high on the list here, but other new features are also considered important.

E. Neural Network

The final approach we consider is a neural network architecture. Due to limited computational resources, we restrict ourselves to the simple case of a neural network with 3 linear layers, dropout and ReLu activation. First, tuning the hyperparameters with 10 trials (the details here are omitted for brevity due to the complexity of the model), we identify

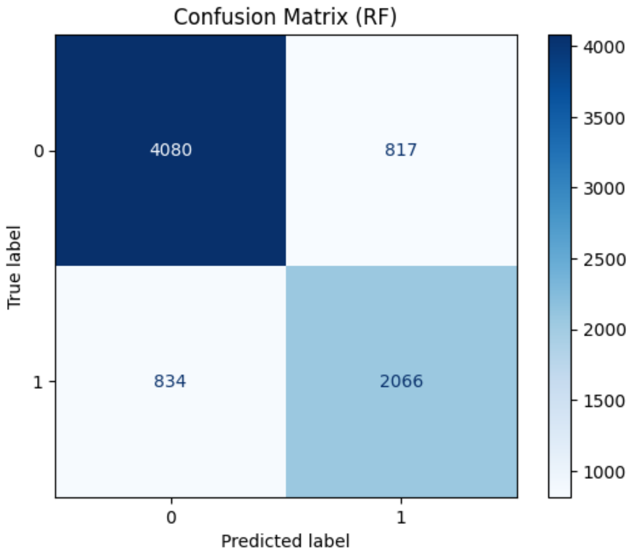

Fig. 6. Confusion Matrix for Random Forest

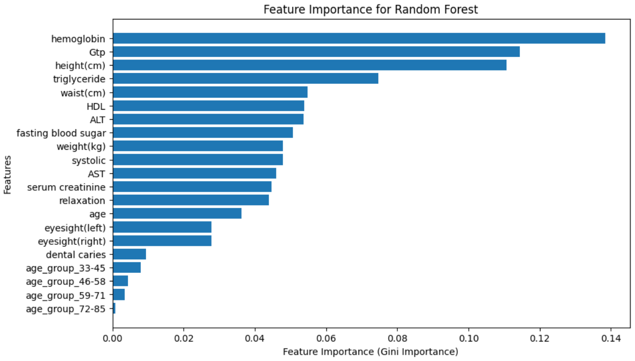

Fig. 7. Feature Importance Plot for Random Forest

the best hyperparameters to be a learning rate of 0.00187, a batch size of 64, a dropout rate of 0.478, and 23 learning epochs. With these parameters, we achieve a cross-validated accuracy score of 0.7533666794921123. Fitting this model and testing it using our test set, we achieve an accuracy of 0.71 with the ROC curve (Figure 12). (For brevity, these figures will be shown in Appendix A.) Note the AUC score of 0.79, which, although lower than the other approaches, is still a reasonable score. To further analyze the accuracy, we consider the confusion matrix (Figure 13). The model has more of both FPs and FNs, particularly FNs, with a rate of 0.16 (comparable to the base logistic regression ratea). We consider the SHAP plot (computed only on 100 samples due to limited computational resources, shown in Figure 14). Here, the most impactful predictors are Gtp, hemoglobin, and height.

V. Discussion

Overall, all models performed reasonably well, with the Random Forest approach standing out as the best by a fairly wide margin. Each of our models likely generalizes well to new data, given that the cross-validated predicted accuracy was close to the real accuracy on the test set. While the Random Forest performs the best, it is also the most computationally expensive model to explain. Of the other, worse-performing models, Linear Regression is the simplest and cheapest to run, and so too has its place in an applied setting (perhaps in the case of very limited available hardware, such as in impoverished regions).

The SVM and Neural Network approaches, on the other hand, were fairly unsuccessful. In the SVM case, it may be that the problem is not well-suited to the approach. However, our implementation of the NN approach was severely restricted by our limited access to computational resources. A more complex model could well outperform even the RF approach and merits further study. Other potential avenues for improvement to all models include setting a loss function that punishes FNs more harshly, to incite our models to evade FNs as much as possible.

Although the Random Forest model achieves a quite low FN rate of 0.11, this is still not nearly low enough to rely entirely on the model, given the life-and-death nature of the application. As such, in the practical setting, this model is best used in conjunction with other methods of determining the smoking status. Since false positives are not dangerous to the patient, a physician may trust a positive prediction of smoking status as given, but given a negative prediction, it should be reinforced by interviewing the patient.

Based on the SHAP (or feature importance plot) for each model, we note that indicators of liver and kidney damage are consistently strong predictors for smoking status. We may conclude that there is a correlation between smoking and damage to these organs. The Random Forest model in particular also highlights hemoglobin levels in the blood as well as height and weight: height and weight in particular may indicate a bias towards male smoker status as opposed to female, given there is no biological reason to expect these measurements to indicate the smoker status directly. Overall, this unexpected result may merit further study.

VI. Conclusion

We have explored several machine learning approaches to predict smoker status based on patient bio-signals. Overall, the best-performing model was a Random Forest based on 100 estimators, which achieved an accuracy of 0.79, an AUC score of 0.87, and a false-negative rate of 0.11: particularly important due to the damage potentially caused by a false belief that a patient is a non-smoker. The important features for the prediction of smoking status shared between all approaches tested appear to be common indicators of liver and kidney damage. The least computationally expensive model, with a still-reasonable accuracy of 0.72 and false negative rate of 0.16, is simple Logistic Regression, which may also potentially have applications in practical applications.

VII. Appendix A: Extra Plots

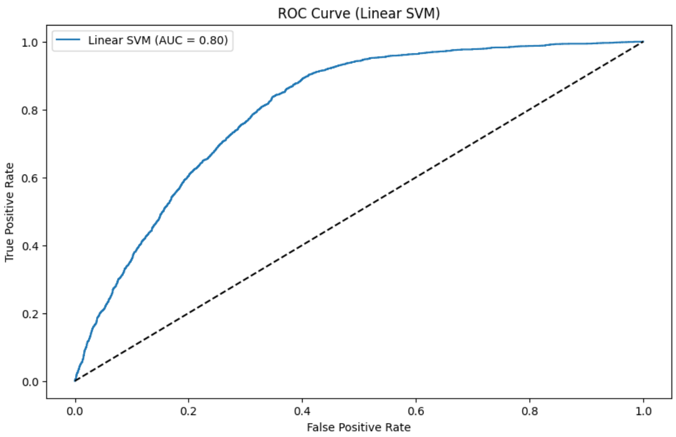

Fig. 8. ROC Curve for Linear SVM

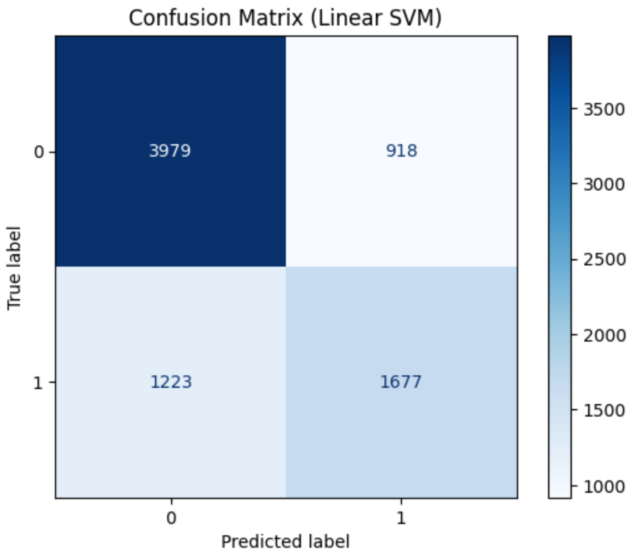

Fig. 9. Confusion Matrix for Linear SVM

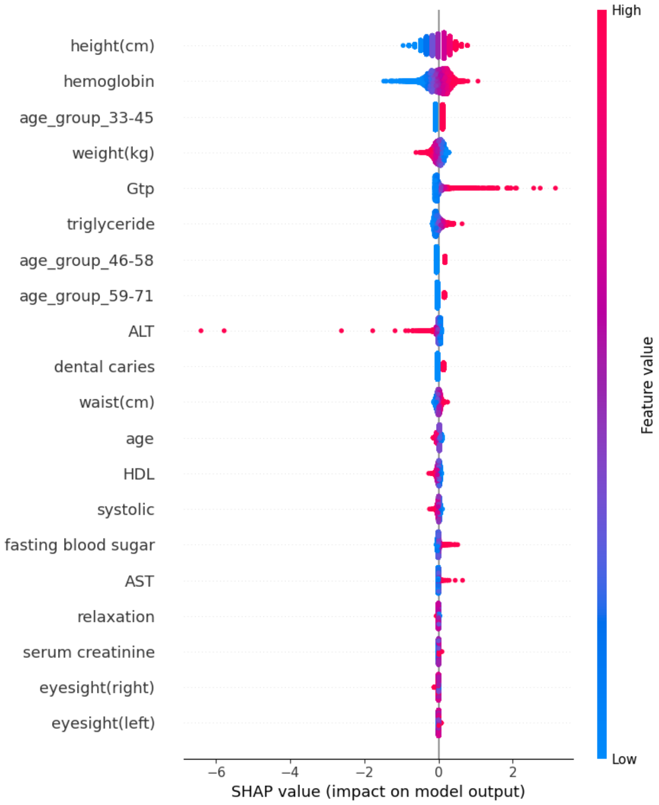

Fig. 10. SHAP Plot for Linear SVM

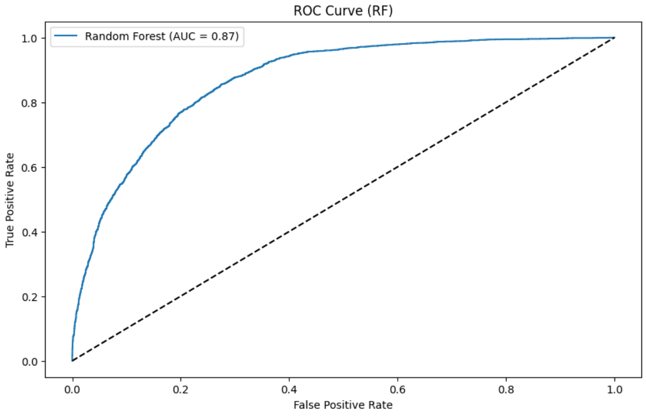

Fig. 11. ROC Curve for Random Forest

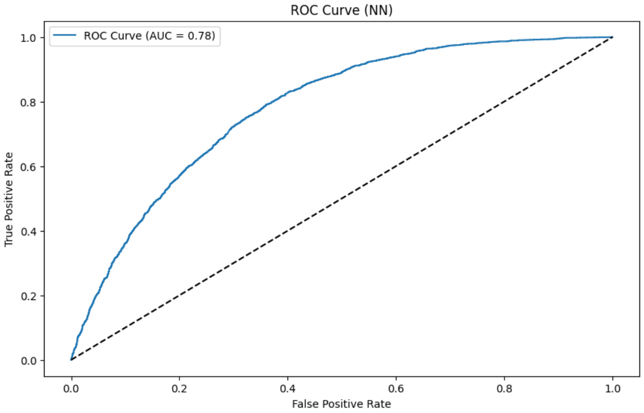

Fig. 12. ROC Curve for Neural Network

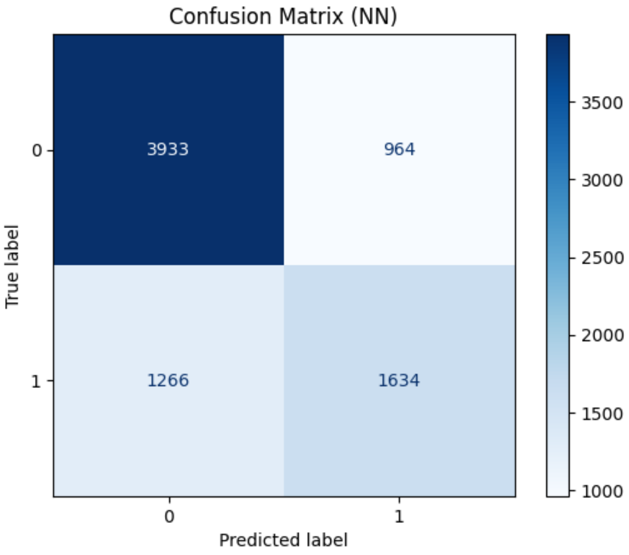

Fig. 13. Confusion Matrix for Neural Network

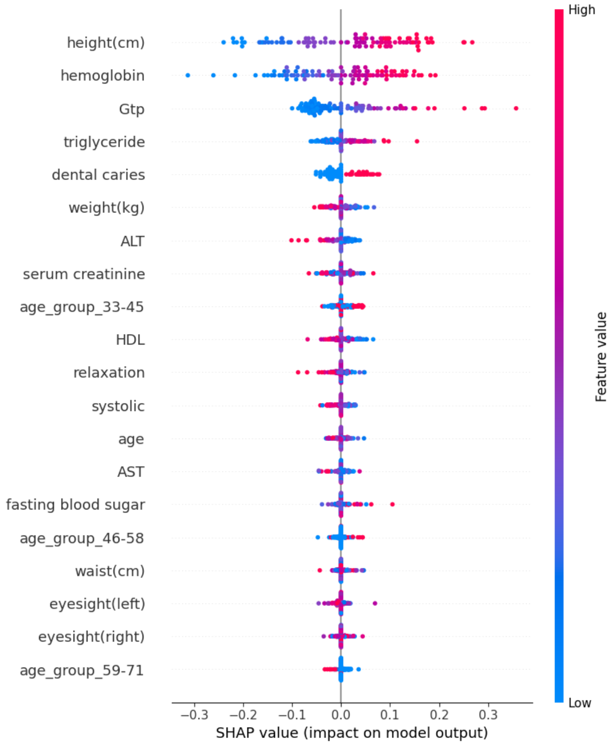

Fig. 14. SHAP Plot for Neural Network (Small Sample)

VIII. Appendix B: Contributions and Code Availability

All members contributed equally to the conceptualization of the project, with each team member taking the lead on a specific portion while receiving equal support from the rest of the team. Jonathan Oxman led the report drafting process, with the team providing feedback, proofreading, and aligning on the trajectory for the coding and presentation focus. Jie Min and Yifu Zheng took the initiative in developing the coding portion of the project, with active participation from all members. Huilin Niu and Muxuan Sun organized the presentation details, and the team worked collaboratively to design the layout and aesthetics for result visualizations. We gratefully acknowledge the contributions of every team member and extend our appreciation to the DATASCI course teaching team for their guidance and support, which ensured the success and final output of this project.

The code can be found at https://github.com/JeffZheng021/BinarySmoke/blob/main/.

References

[1] U.S. Department of Health, Centers for Disease Control Human Services, National Center for Chronic Disease Prevention Prevention, Office on Smoking Health Promotion, and Health. The health consequences of smoking—50 years of progress. Technical report, U.S. Surgeon General, 2014.

[2] A. Mishra, P. Chaturvedi, S. Datta, S. Sinukumar, P. Joshi, and A. Garg. Harmful efffects of nicotine. *Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology*, 2015.

[3] M.A. Carrick, J.M. Robson, and C. Thomas. Smoking and anaesthesia. *BJA Education*, 19(1):1–6, January 2019.

[4] R Murray, J Leonardi-Bee, J Marsh, L Jayes, and J Britton. Smoking status ascertainment and interventions in acute medical patients. *Clinical Medicine*, 12(1):59–62, February 2012.

[5] L Lores Obradors, E Monsó Molas, A Rosell Gratacós, I Badorrey, and I Sampablo Lauro. Do patients lie about smoking during follow-up in the respiratory medicine clinic? *Arch. Bronconeumol.*, 35(5):219–222, May 1999.

[6] Smoker Status Prediction using Bio-Signals.