Binary Prediction of Smoker Status using Bio-Signals

Martyna Leśniak, Michał Matuszyk, Hubert Sobociński

June 10, 2024

Abstract

This report presents the development and evaluation of a machine learning model for predicting smoker status based on bio-signals. Smoking significantly impacts health, leading to various diseases and reduced life expectancy, making early detection crucial. Traditional identification methods often rely on self-reported data, which can be unreliable.

Our study utilized a diverse set of physiological and biochemical indicators, including vital signs, blood test results, and other health metrics. After preprocessing and feature extraction, we trained and evaluated multiple machine learning algorithms. The best results were achieved by the Random Forest classifier, which demonstrated high accuracy and robustness.

This reliable, non-invasive method for detecting smoker status can support public health campaigns by providing clear evidence of smoking's health impacts and facilitating early detection of related diseases. Such an approach can help reduce smoking rates and promote healthier lifestyles.

1 Introduction

Smoking is a major health concern worldwide, recognized as a significant risk factor for numerous diseases, including cancer, cardiovascular disorders, respiratory ailments, and reproductive problems. It affects nearly every organ in the body and can reduce life expectancy by about 15 years. As of 2018, smoking is the leading cause of preventable illness and death globally, with the World Health Organization estimating that smoking-related deaths will reach 10 million annually by 2030.

Despite available treatments to help people quit smoking, less than one-third succeed. Many doctors find smoking cessation counseling ineffective and time-consuming, leading to its infrequent use. Traditional methods to identify smokers often rely on self-reported data, which can be unreliable.

To address these issues, we developed a machine learning model to predict whether an individual is a smoker using physiological and biochemical indicators. The dataset comprises two parts: the original data from https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals and additional data generated from a deep learning model trained on the same dataset. The feature distributions in the generated data are similar to the original.

Our goal was to create a machine learning model for predicting smoking status using an expanded dataset to improve accuracy and reliability. Input features included parameters easily collected during a patient's initial visit, making the model practical for both physicians and patients. The model's output would indicate whether an individual is a smoker. This approach ensures the data is easily obtainable, supports practical use in clinical settings, and aids public health campaigns by providing clear evidence of health impacts and facilitating early detection of related diseases.

2 Data preparation

In this section, we describe the steps taken to prepare the data for training and evaluating our machine learning models.

2.1 Exploratory Data Analysis(EDA)

The first step we took to better understand the data, identify patterns, and summarize the main characteristics was performing Exploratory Data Analysis (EDA). We began by getting an overview of our dataset.

```
Data columns (total 23 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   age                  198240 non-null  int64
 1   height(cm)           198240 non-null  int64
 2   weight(kg)           198240 non-null  int64
 3   waist(cm)            198240 non-null  float64
 4   eyesight(left)       198240 non-null  float64
 5   eyesight(right)      198240 non-null  float64
 6   hearing(left)        198240 non-null  int64
 7   hearing(right)       198240 non-null  int64
 8   systolic             198240 non-null  int64
 9   relaxation           198240 non-null  int64
 10  fasting blood sugar  198240 non-null  int64
 11  Cholesterol          198240 non-null  int64
 12  triglyceride         198240 non-null  int64
 13  HDL                  198240 non-null  int64
 14  LDL                  198240 non-null  int64
 15  hemoglobin           198240 non-null  float64
 16  Urine protein        198240 non-null  int64
 17  serum creatinine     198240 non-null  float64
 18  AST                  198240 non-null  int64
 19  ALT                  198240 non-null  int64
 20  Gtp                  198240 non-null  int64
 21  dental caries        198240 non-null  int64
 22  smoking              198240 non-null  int64
dtypes: float64(5), int64(18)
```

Figure 1: Dataset columns info

The dataset included 23 columns, all with numeric values, and there were no missing values. We also checked for duplicated rows and some of them were found.
The next step was visualizing histograms of the features to understand their distributions.

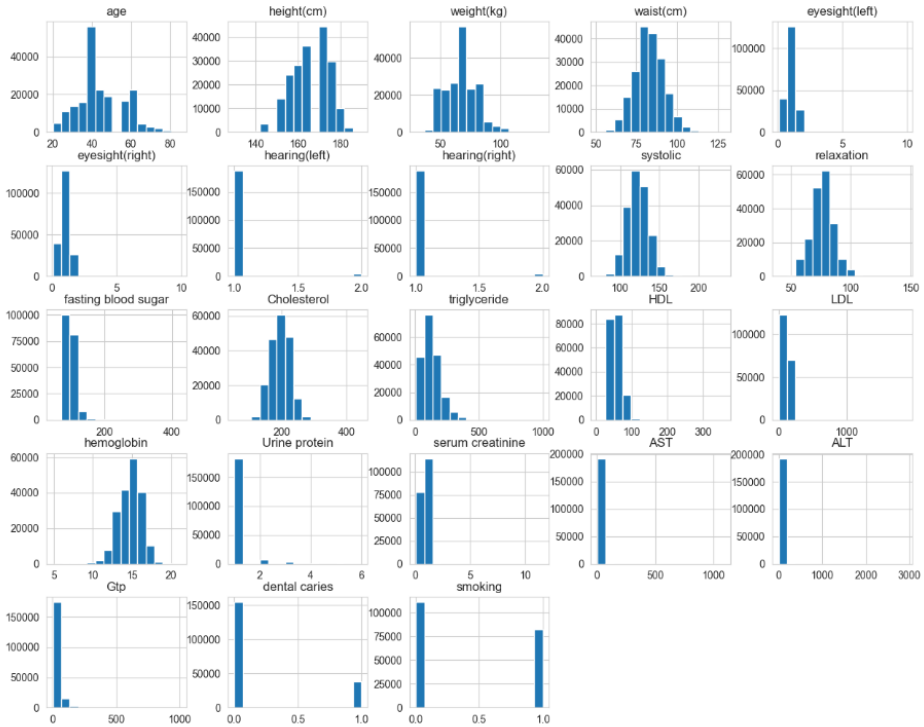

Figure 2: Histograms of features

Based on the histograms and the number of unique values, we observed that the features 'hearing', 'urine protein,' and 'dental caries' were discrete, while the remaining features were continuous. 'Hearing' and 'dental caries' were already binary, and 'urine protein' was ordinal.

We also examined boxplots to identify potential outliers in the data.

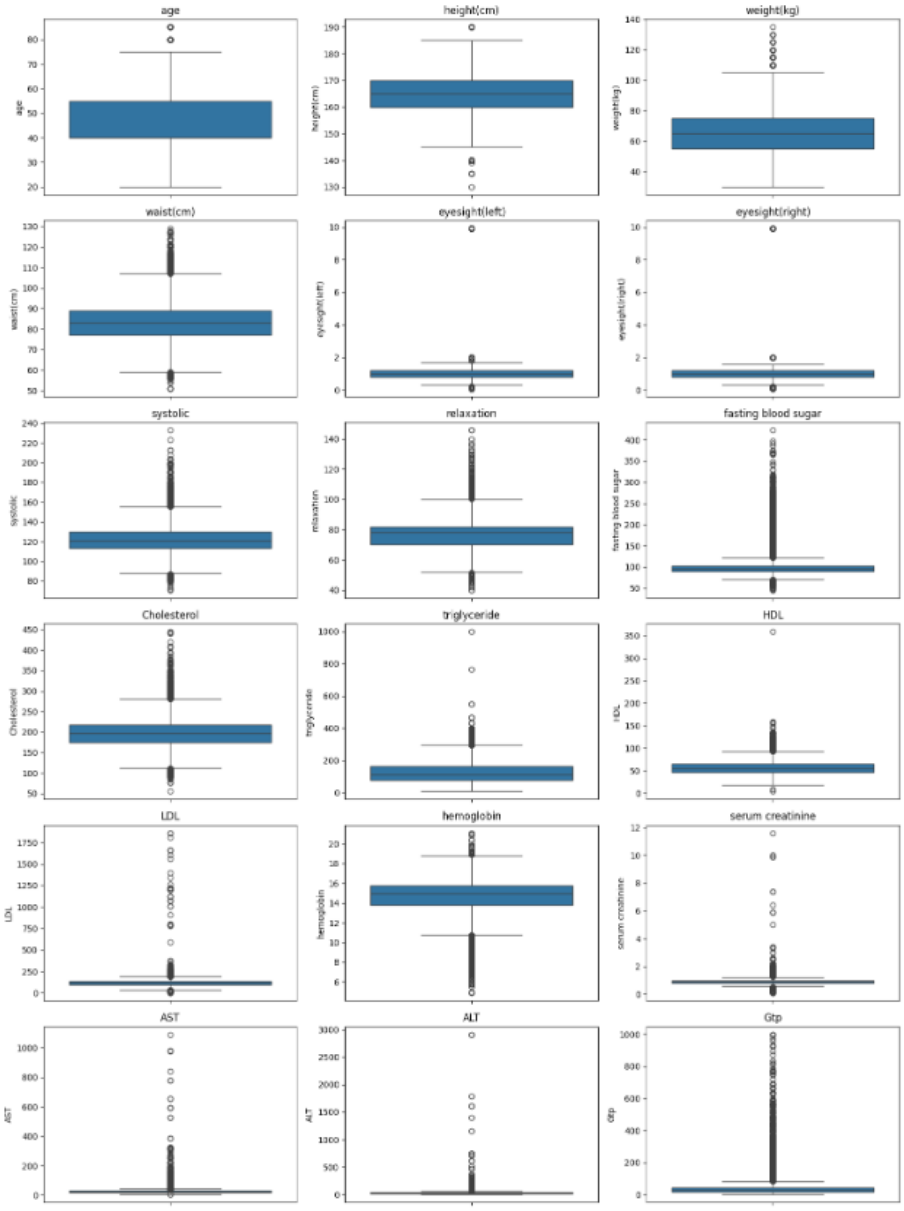

Figure 3: Distribution and of Continuous Features via Boxplots

There were quite a few outliers in our data, and the distribution was sometimes skewed.

Further in our analysis, we checked the boxplots divided by smoking and non-smoking status to identify any potential associations. The only features that showed significant differences based on smoking status were hemoglobin, triglycerides, weight, and age. Additionally, the histogram for dental caries indicated noticeable differences.

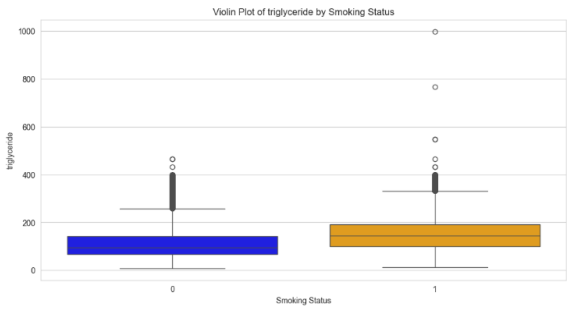

Figure 5: Distribution of triglyceride by smoking status

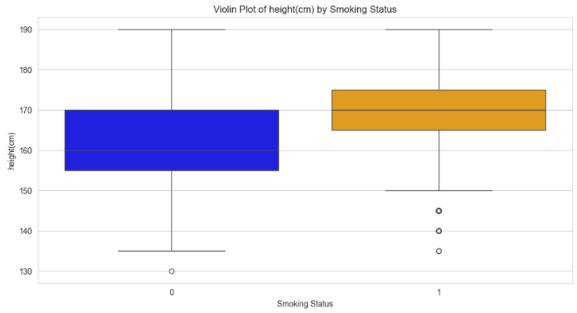

Figure 4: Distribution of height by smoking status

[H]

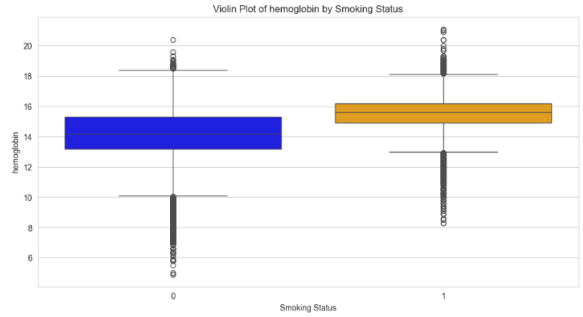

Figure 6: Distribution of hemoglobin by smoking status

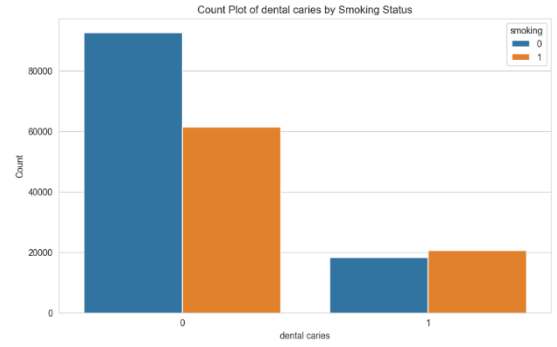

Figure 7: Occurrence of dental caries by smoking status

To conclude this stage, we examined the relationships between the individual parameters and our target variable using a correlation matrix.

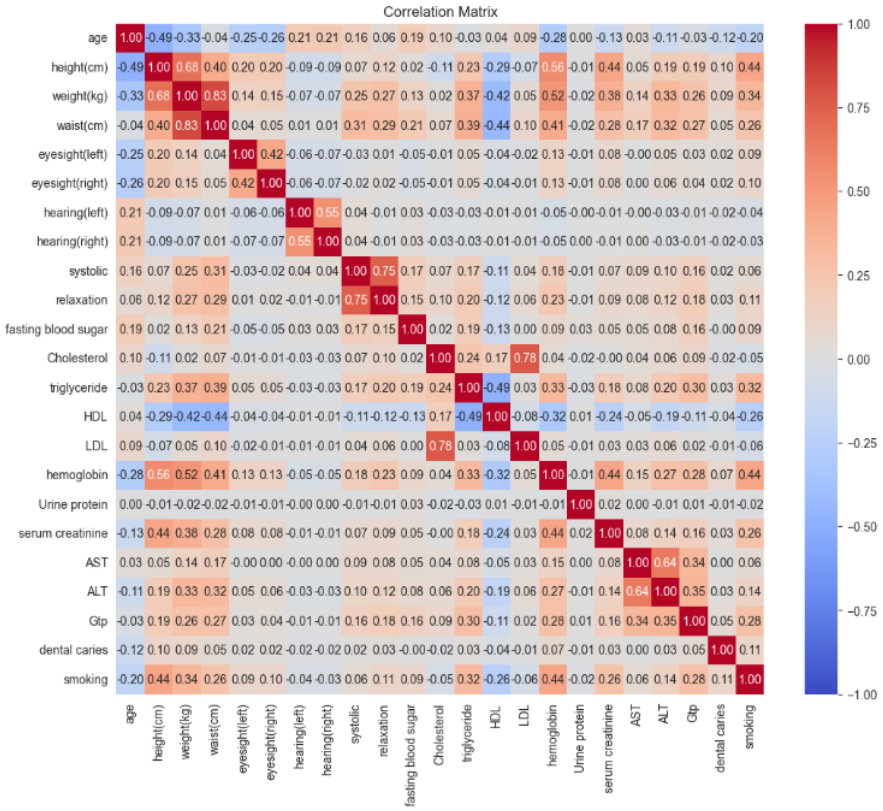

Figure 8: Correlation matrix

From the correlation matrix, we observed that the features with the highest correlations are weight and waist, eyesight (left) and eyesight (right), hearing (left) and hearing (right), LDL and cholesterol, and AST and ALT. The features that showed the strongest correlation with the target variable (smoking status) are height, triglycerides, and hemoglobin. It can be suspected that height is mainly related to gender.

2.2 Data Preprocessing

Based on our EDA, we took the following steps to preprocess the data for modeling:

1. **Removing Duplicates:** We checked for duplicated rows and, since duplicates were found, we removed them to ensure data integrity.

2. **Handling outliers**

   - **Visual Acuity Outliers:** The range for visual acuity is between [0.1, 2.0]. Values exceeding this range are likely errors. We addressed this by setting any 'eyesight(left)' or 'eyesight(right)' values greater than 2.0 to the median of their respective columns.
   - **Other Outliers:** Features such as 'triglyceride', 'HDL', 'AST', and 'ALT' exhibited a large spread in their boxplots. Therefore, we capped the outliers by using the 99th percentile as the upper limit and the 1st percentile as the lower limit. Values above the upper limit were set to the 99th percentile value, and values below the lower limit were set to the 1st percentile value.

3. **Log Transformation:** To normalize the distribution of certain skewed features, we applied a log transformation. The features transformed were 'systolic', 'relaxation', 'fasting blood sugar',

'Cholesterol', 'triglyceride', 'HDL', 'LDL', 'serum creatinine', 'AST', 'ALT', and 'Gtp'. The log transformation was performed by taking the natural logarithm of each value plus one.

4. **Creating New Columns:** To enhance the dataset, we created new columns:
   - **BMI:** Body Mass Index (BMI) calculated using the formula $\text{BMI} = \frac{\text{weight(kg)}}{(\text{height(cm)}/100)^2}$.
   - **Hearing:** A composite hearing column created based on the values of 'hearing(left)' and 'hearing(right)':
     - 1 if both ears have a hearing value of 1.
     - 2 if one ear has a hearing value of 1 and the other ear has a hearing value of 2.
     - 3 if both ears have a hearing value of 2.
   - **Eyesight:** A composite eyesight column created by averaging 'eyesight(left)' and 'eyesight(right)'.
   - **AST/ALT Ratio:** The ratio of AST to ALT.
5. **Removing Unnecessary Columns:** We removed the 'hearing(left)', 'hearing(right)', 'eyesight(left)', 'eyesight(right)', and 'Cholesterol' columns for the following reasons:
   - **Hearing and Eyesight:** These columns were not necessary because we created composite 'hearing' and 'eyesight' columns, which were highly correlated with the original columns.
   - **Cholesterol:** This column was highly correlated with 'LDL', which had a stronger correlation with the target variable. Additionally, its correlation with other features was similar to 'Cholesterol', making it redundant.

   We decided to retain the other columns and defer further feature selection to a later stage.
6. **Normalization:** The final step was normalizing the features using Min-Max scaling so all features were scaled to a range between 0 and 1. This method was chosen because the distribution of most features was not normal, making Min-Max scaling more appropriate for bringing all features into a common scale without distorting their underlying distributions.

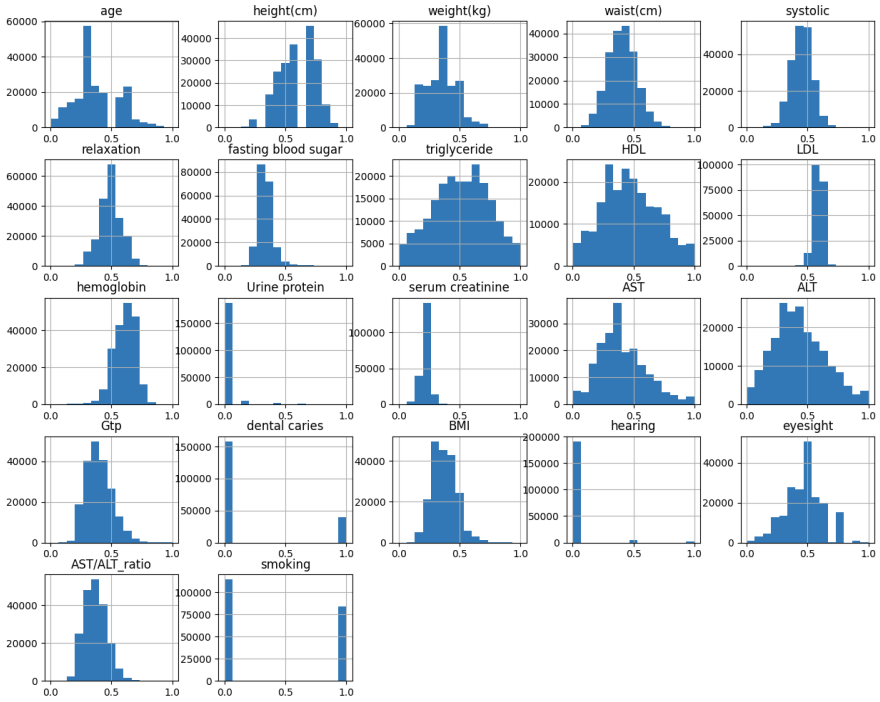

Figure 9: Data after preprocessing

3 Preliminary Modeling

After preprocessing the data, we proceeded with preliminary modeling to evaluate the performance of various algorithms and select the best-performing ones for further optimization.

The dataset was first split into training (80%) and testing (20%) dataset.

The models that we selected for initial evaluation include:

- Decision Tree
- Support Vector Machine (SVM)
- Logistic Regression
- K-Neighbors
- Naive Bayes
- Random Forest
- Neural Networks

Each model was trained on the training set using default parameters and evaluated on the testing set. The evaluation metrics included accuracy, precision, recall, F1-score, and ROC AUC score. Our primary goal was to accurately identify smokers, making recall the most critical metric for our purposes. Additionally, we aimed to achieve the highest possible ROC AUC score.
The results are as follows:

ML model	**class**	**precision**	**recall**	**f1-score**	**ROC AUC**
Decision Tree	0	0.74	0.75	0.74	0.6941
	1	0.65	0.64	0.65	0.6941
Support Vector Machine (SVM)	0	0.66	0.47	0.55	0.7905
	1	0.49	0.67	0.56	0.7905
Logistic Regression	0	0.79	0.77	0.78	0.8344
	1	0.7	0.72	0.71	0.8344
K-Neighbors	0	0.77	0.75	0.76	0.7988
	1	0.68	0.7	0.69	0.7988
Naive Bayes	0	0.8	0.68	0.74	0.7905
	1	0.64	0.77	0.7	0.7905
Neural Network	0	0.82	0.75	0.78	0.8500
	1	0.7	0.78	0.74	0.8500
Random Forest	0	0.84	0.77	0.8	0.8635
	1	0.72	0.8	0.76	0.8635

Based on these results we decided to pursue to improvement of the Neural Network and Random Forest.

4 Model Optimization and Enhancement

4.1 Neural Network

Firstly we made improvements to the Neural Network, by changing:

- Number of hidden layers: 1-4
- Number of neurons in each layer: 1-22
- Changing the activation function: relu, sigmoid
- Number of epochs: 10, 20, 50, 100

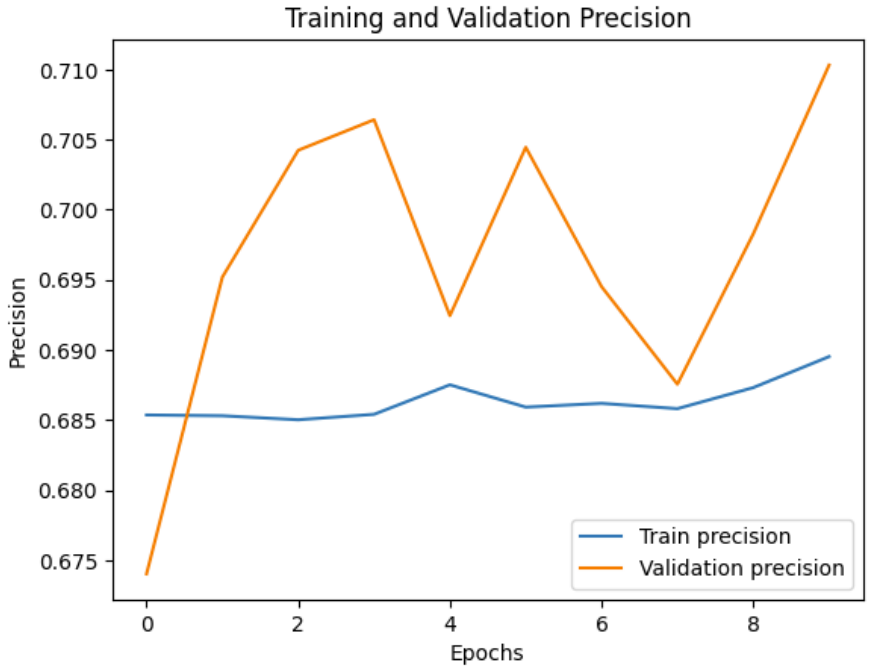

Figure 10: Precision over epochs for the Nueral Network

Some sample results are shown below:

The largest gain happened when the data used was preprocessed as opposed to not - the ROC-AUC score went up from 0.75 to over 0.84. The best result was seen by the following neural network:

- **Input Layer**: The network starts with a dense layer comprising 64 neurons, which serves as the input layer. This layer utilizes ReLU (Rectified Linear Unit) as its activation function. The input dimensionality (`input_dim`) is set to the number of features in the training dataset (`X_train.shape[1]`), ensuring that the model can process the feature set directly.
- **Batch Normalization and Dropout**: Following the first dense layer, a batch normalization layer is employed. This layer normalizes the activations from the previous layer at each batch, stabilizing the learning process and reducing the number of epochs required to train the model. After normalization, a dropout layer with a dropout rate of 0.3 is added. This dropout layer randomly sets a proportion of input units to zero during training, which helps prevent overfitting by reducing co-dependence among neurons.
- **Hidden Layers**:
  - The second layer of the network is another dense layer with 32 neurons, also employing the ReLU activation function. This is followed by another batch normalization and a dropout layer with the same dropout rate of 0.3.
  - Subsequently, the model includes a third dense layer containing 16 neurons, using ReLU activation. This layer is similarly followed by batch normalization and dropout layers, with the dropout again set at 0.3.
- **Output Layer**: The final layer of the model is a dense layer with a single neuron, using the sigmoid activation function. This output layer is designed to produce a probability indicating the likelihood that the given input belongs to the positive class, appropriate for the binary classification task at hand.

The results are as follows:

However the Neural Network had a tendency to under classify - as there where fewer smokers and having chosen Recall as the metric and deciding that the precision of smokers was of the utmost importance, we decided to change the model to Random Forest

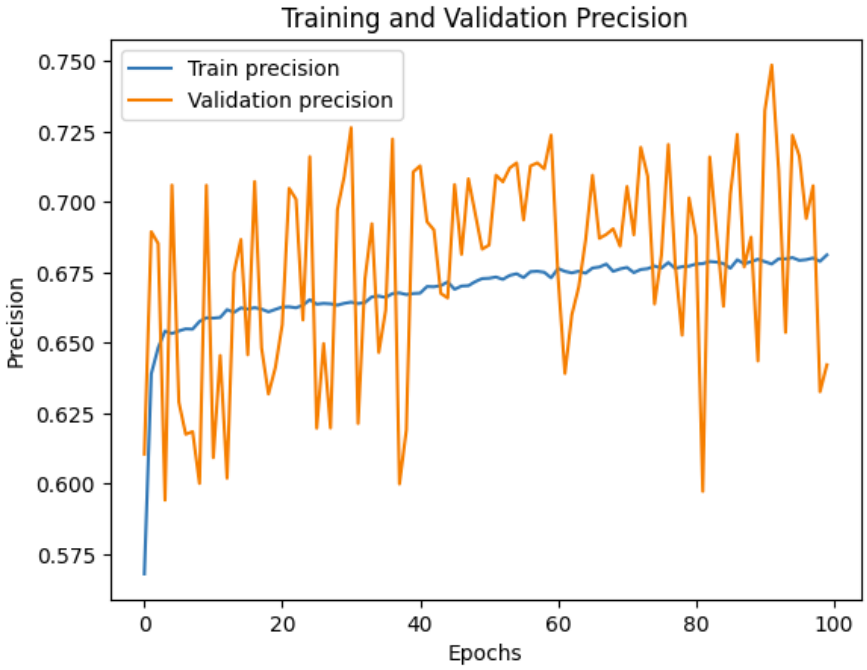

Figure 11: Best Neural Network model precision over the epochs

Class	Precision	Recall	F1-score	Support
0	0.89	0.60	0.72	17783
1	0.64	0.91	0.75	14069
Accuracy				
0.74				
Macro Average				
Precision: 0.77, Recall: 0.75, F1-score: 0.73				
Weighted Average				
Precision: 0.78, Recall: 0.74, F1-score: 0.73				

Table 1: Classification Report

4.2 Best model - Random Forest

The best results were achieved by the Random Forest classifier. We focused further efforts on tuning the Random Forest classifier by searching for the best hyperparameters using Grid Search.

```
Random Forest Model
Accuracy: 0.7806698950766747
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.78      0.80     22677
           1       0.72      0.79      0.75     16971

    accuracy                           0.78     39648
   macro avg       0.78      0.78      0.78     39648
weighted avg       0.78      0.78      0.78     39648

--------------------------------------------------
```

Figure 12: Classification report of Random Forest Classifier with default parameters

After identifying the optimal hyperparameters, the model's performance showed some improvement.

The parameters found using GridSearch:

bootstrap=False
max_depth=None
max_features='sqrt'
min_samples_leaf=2
min_samples_split=5
n_estimators=300
as demonstrated in the following results:

```
Random Forest Model
Accuracy: 0.7823345439870864
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.77      0.80     22677
           1       0.72      0.80      0.76     16971

    accuracy                           0.78     39648
   macro avg       0.78      0.78      0.78     39648
weighted avg       0.79      0.78      0.78     39648

--------------------------------------------------
```

Figure 13: Classification report of Random Forest classifier with optimized hyperparameters

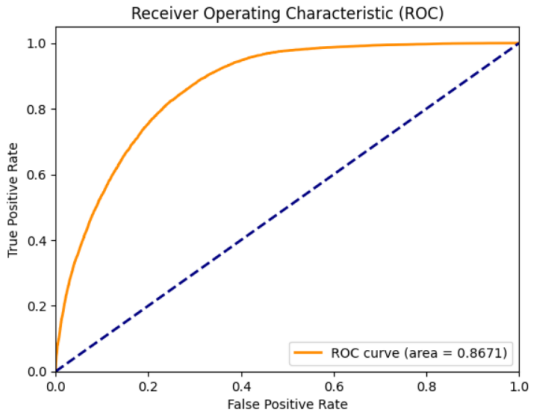

Figure 14: ROC curve

The recall for smokers reached 80%, which indicates that the model effectively identifies the majority of actual smokers, reducing the number of false negatives.

We found that height, hemoglobin content and Gtp had the greatest importance, which is shown in this plot:

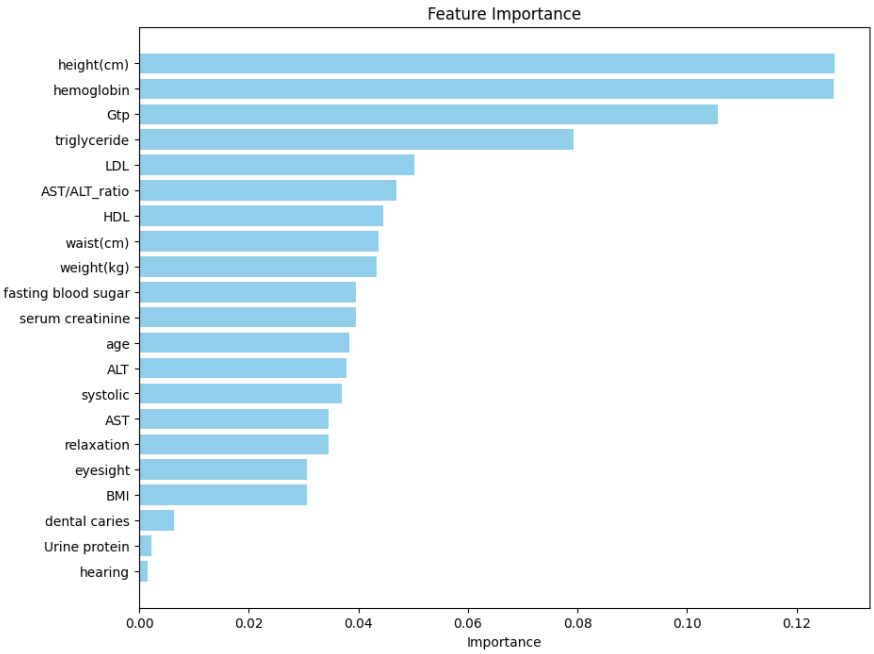

Figure 15: Feature importances plot

After investigating the connection between hemoglobin levels and smoking, we found out that research from 2017 states, "smokers have significantly higher levels of hemoglobin"[1] .This correlation is due to the body's response to chronic carbon monoxide (CO) exposure from smoking. CO binds to hemoglobin more effectively than oxygen, reducing the blood's oxygen-carrying capacity. To compensate, the body increases hemoglobin production, ensuring sufficient oxygen transport despite the presence of CO.

The second feature, height, is probably correlated to sex. According to the research, "smoking prevalence in men is more than four times that in women globally, but the difference is much less in most parts of Europe, and Eastern Europe as a whole has the highest smoking prevalence of any region in the world.[2]"

5 Competition

Our performance in the recent competition is summarized in the table below. This table presents our private scores for various submissions:

1	0.80631
2	0.84169
3	0.86006
4	0.85931
5	0.86013
6	0.86004
7	0.86154
8	0.85381
9	0.86116
10	0.85784
11	0.85916
12	0.85916
13	0.85916

As illustrated, our models achieved a range of scores, with the highest being 0.86154 and the lowest 0.80631. These scores indicate consistent performance across different models and configurations.

Furthermore, the top 10 participants in the competition and their respective scores are listed in the following table:

Position	Username	Score
1	Pearl-Luck	0.87946
2	Chan Lee	0.87944
3	Ravi Ramakrishnan	0.87938
4	aldparis	0.87935
5	kaggle_JJ	0.87934
6	DJ_C	0.87932
7	Sarun P M	0.87926
8	Master Jiraiya	0.87922
9	baellouf	0.87921
10	HyukJunCho	0.8792

Our models' performances were within 1-2% of the top position on the leaderboard. This indicates a high level of competitiveness and the effectiveness of our approaches. The narrow margin suggests that with further fine-tuning and optimization, it is plausible to reach or surpass the highest scores achieved in the competition, however due to the limited nature of our machines, we weren't able to perform the level of fine-tuning we wanted to.

6 Conclusion

The findings indicate that bio-signals can effectively distinguish smokers from non-smokers, offering a reliable, non-invasive detection method. This approach can be instrumental in public health campaigns, providing clear evidence of smoking's health impact and aiding early detection of related diseases. By highlighting health risks and enabling early intervention, this method supports efforts to reduce smoking rates and promote healthier lifestyles. Future work will focus on expanding the dataset, refining the model, and exploring clinical applications.

References

[1] "Effect of Cigarette Smoking on Haematological Parameters in Healthy Population." Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5511531/

[2] "Tobacco smoking: Health impact, prevalence, correlates and interventions." Available: https://www.tandfonline.com/doi/full/10.1080/08870446.2017.1325890