International Journal of Computing and Digital Systems
2025, VOL. 17, NO. 1, 1–14

http://dx.doi.org/10.12785/ijcds/1571016631

Biosignals Based Smoking Status Prediction Using Standard Autoencoder And Artificial Neural Network

Dr. N. Palanivel[1], **S. Deivanai**[2], **G. Lakshmi Priya**[2] **and B. Sindhuja**[2]

[1]*Associate Professor, Dept. of Computer Science and Engineering, Manakula Vinayagar Institute of Technology, Puducherry, India*
[2]*UG Students, Dept. of Computer Science and Engineering, Manakula Vinayagar Institute of Technology, Puducherry, India*

Received 5 April 2024, Revised 14 September 2024, Accepted 19 September 2024

Abstract: Smoking has emerged as a critical global health issue, leading to numerous diseases and premature mortality across the world. Utilizing biosignals to predict smoking status can yield insightful information for tailored interventions and smoking cessation programs. This work presents a novel method that combines an artificial neural network (ANN) and a regular autoencoder to predict smoking status based on biosignals. The proposed method involves preprocessing biosignal data to extract relevant features, which are then input into an autoencoder for dimensionality reduction. The output of Standard AE is given as input to ANN for smoking status prediction. This hybrid model is trained and tested on a dataset comprising biosignal data from individuals using supervised learning. The suggested strategy is effective, as seen by the experimental results, which show a high degree of prediction accuracy about smoking status. The model's performance is further validated through comparisons with existing methods, showing superior performance in terms of accuracy and robustness. The proposed model is integrated into a user-friendly application, along with specific web pages focused on informing users about the harmful effects of smoking and the benefits of quitting. The platform provides users with personalized insights regarding their smoking status based on biosignals. The implemented deep learning model achieves the desired level of accuracy in predicting smoker status, and the user-friendly application offers a convenient platform for public health and personalized healthcare interventions.

Keywords: Biosignals, Smoking Prediction, Standard Autoencoder (Standard AE), Artificial Neural Network (ANN), Dimensionality Reduction, Smoking Cessation.

1. INTRODUCTION

Smoking remains one of the most pressing public fitness-traumatic situations globally, exacting a heavy toll on people's health and well-being and straining healthcare systems. In spite of worldwide reputation campaigns and crook responsibilities, smoking remains a wonderful contributor to avoidable deaths and ailments. Predicting smoking reputation correctly is essential for developing powerful smoking cessation strategies and personalized interventions aimed at decreasing smoking incidence and its associated fitness burdens.

In current years, there has been a developing hobby of leveraging biosignals to count on smoking fame. Biosignals, inclusive of physiological and biochemical markers obtained from people, offer precious insights into their health popularity and behaviours, along with smoking behavior. Robust predictive models that reliably affirm a person's smoking fame may be advanced by using state-of-the-art computer techniques to investigate those biosignals.

This paper gives a completely unique approach for predicting smoking reputation based totally on biosignals using a fusion of autoencoder and artificial neural community (ANN) architectures. Autoencoders are unsupervised deep learning algorithms able to learn inexperienced representations of input records with the useful resource of encoding them right into a decrease-dimensional latent area and then reconstructing them. By leveraging the latent representations found by an autoencoder, in conjunction with additional features extracted from biosignal data, an ANN version is educated to expect smoking recognition appropriately.

The proposed technique has several advantages over conventional strategies of smoking status prediction. Firstly, it allows the automatic extraction of applicable capabilities from biosignal records, putting off the need for manual characteristic engineering and doubtlessly taking pictures of diffused patterns indicative of smoking behavior. Secondly, the fusion of autoencoder and ANN architectures

E-mail address: palanivelcse@mvit.edu.in, deivanaisubramanian6@gmail.com

allows the model to investigate complicated relationships between biosignals and smoking, improving its predictive performance.

Extending a predictive version that can accurately determine smoking popularity based on biosignal facts is the main goal of this research. In order to gain from this, the actions listed below are taken:

- Data Preprocessing and Feature Extraction: Biosignal records, along with physiological and biochemical parameters including age, weight, levels of cholesterol, and so forth., are pre-processed to ensure consistency and reliability. Relevant functions are extracted from the biosignal facts, together with the uncooked measurements and derived metrics.

- Autoencoder is primarily based on Dimensionality Reduction: The pre-processed biosignal information is entered into an autoencoder, which learns a lower-dimensional representation of the entered facts. This latent instance captures the important capabilities of the biosignals while decreasing noise and, besides the factor information.

- ANN is primarily based on Smoking Status Prediction: The latent representations observed through the usage of the autoencoder, collectively with additional features, are used as input for training an ANN model to assume smoking reputation. The ANN version is professional in the usage of supervised studying strategies, in which the goal variable is the binary classification of smoking repute (smoker or non-smoker).

- Model Evaluation and Validation: The mean efficacy of the developed prediction model is assessed by a variety of criteria, including F1-score, accuracy, precision, and consideration. In order to assess the model's generalization functionality, flow validation techniques are also used during testing.

- Integration into a user-friendly application: The advanced predictive model is included in consumer-pleasant software aimed at promoting smoking cessation. The application offers users custom-designed insights into their smoking recognition primarily based on biosignals, along with educational assets on the detrimental consequences of smoking and the blessings of quitting. A menu-frequently-based chatbot is blanketed to deal with user queries and offer useful resources for smoking cessation efforts.

Overall, this research contributes to the improvement of predictive modelling techniques for smoking prediction based on biosignals. The advanced version gives a promising technique for identifying individuals who smoke and non-people who smoke appropriately, thereby facilitating focused interventions and public health responsibilities geared toward decreasing smoking incidence and improving population health consequences.

2. RELATED WORKS

Machine learning applications for predicting smoking behaviour conducted by Etter et al. [1] explores the domain of smoking cessation interventions, with a specific focus on predicting changes in smoking behavior six months after utilizing the Stop-Tabac smartphone app. Drawing upon data from a randomized trial, the researchers utilized machine learning techniques, particularly the Random Forest classification algorithm, to discern predictors associated with smoking cessation, reduction, and relapse. The study unveils the intricate interplay of factors influencing smoking behaviour, ranging from motivational factors to levels of dependence and the utilization of nicotine replacement products. Although the results do not explicitly declare the accuracy of the machine learning models, they do point to promising results when it comes to tailoring smoking cessation programs to the unique profiles of individual users. This study provides information that may improve the effectiveness of apps designed to help people quit smoking and, in turn, improve public health outcomes related to the battle against tobacco use. In the area of cessation applications, it represents a substantial breakthrough.

The study conducted by Issabakhsh et al. uses information from the Population Assessment of Tobacco and Health (PATH) survey [2] to investigate the use of machine learning techniques in forecasting adult US smoking cessation. The authors uncover factors that influence smoking cessation and create machine learning classification models targeted at adult smokers who are currently smokers through analysis of waves 1-3 of the PATH project. The study predicts smoking cessation with 72% accuracy by wave 2 and 70% accuracy by wave 3 using algorithms like random forest and gradient boosting machines. The use of e-cigarettes in the past, smoking cigarettes before stopping, age when smoking started, number of years of smoking, use of poly-tobacco, and BMI are among the important factors that were found. This research highlights the significance of machine learning in understanding smoking cessation dynamics and provides insights for the development of effective cessation interventions.

Nuryunarsih et al.'s study demonstrates a significant stride in healthcare analytics by employing artificial neural network (ANN) machine learning [3] to forecast smoking behaviour among Indonesian health professionals. By employing ANN algorithms, the researchers have developed predictive models capable of accurately analysing various factors influencing smoking behaviour, including demographic information, knowledge about smoking-related diseases, and attitudes towards smoking cessation. Achieving an impressive accuracy rate of 81%, these models provide valuable insights into the determinants of smoking behaviour within the Indonesian healthcare workforce. The integration of ANN technology holds promise for enhancing smoking cessation interventions and advancing tobacco control initiatives in Indonesia, ultimately leading to improved patient and public health outcomes. Continued collaboration between healthcare professionals and AI systems is crucial for refining predictive models and ensuring their effectiveness in addressing public health challenges.

Frank et al.'s research paper [4] presents a fresh method to healthcare analytics by predicting smoking status through statistical analysis and machine learning algorithms. By using a range of classification algorithms and statistical approaches, the researchers expect to accurately predict smoking behaviour based on medical information acquired during hospital stays. Using a dataset of 40,000 patients from a community hospital in the Greater Pittsburgh Area, significant differences in blood test results between smokers and nonsmokers are discovered. The INR, HB, and HCT levels are where these variations are most noticeable. The study demonstrates that the logistic model performs better than the other models in terms of accuracy rates when it comes to predicting smoking status. The rates for precision, recall, and F-measure are 83%, 83.4%, 83.2%, and 83.44%, respectively. In conclusion, the current study shows how machine learning algorithms and statistical analysis can predict smoking behavior, improving healthcare analytics and assisting tobacco-reduction-focused public health efforts. In conclusion, the current study shows how machine learning algorithms and statistical analysis can predict smoking behavior, improving healthcare analytics and assisting tobacco-reduction-focused public health efforts. To improve smoking cessation efforts and refine predictive models, data scientists and healthcare practitioners must work together.

Thakur et al. provide a multi-class classification model based on machine learning for the real-time prediction of smoking behavior during activities of daily living (ADLs) [5]. Using data collected from wrist-wearable devices equipped with IMU sensors, the study focused on activities such as walking, running, and smoking. Time-domain, frequency-domain, and descriptive features were retrieved from streaming sensor data using a sliding window method. The developed models achieved high predictive accuracy rates of up to 98.7% for identifying smoking activity, indicating potential for real-time detection and intervention. These findings offer implications for monitoring and motivating smokers to quit, improving healthcare interventions, and expanding into preventive healthcare applications.

Lai et al. concentrate on creating machine learning models to forecast the results of quitting smoking [6]. The research employs a range of machine learning algorithms, such as support vector machines (SVM), artificial neural networks (ANN), logistic regression (LoR), random forests (RF), k-nearest neighbors (KNN), naïve Bayes (NB) and classification and regression trees (CART), and is carried out on patients participating in a smoking cessation program in Northern Taiwan. With a sensitivity of 0.704, specificity of 0.567, accuracy of 0.640, and area under the receiver operating characteristic (ROC) curve of 0.660, the ANN model performed marginally better. By incorporating patient parameters such as nicotine dependence level, daily cigarette consumption, and previous quit attempts, the developed predictive model offers a personalized approach to smoking cessation treatment. These predictive models have the potential to assist healthcare providers in identifying smokers likely to succeed in quitting, thereby enhancing counselling and treatment strategies and ultimately reducing the burden of smoking-related diseases on public health.

Caccamisi et al. Examine how machine learning and natural language processing (NLP) approaches can be combined to automatically extract and classify patients' smoking status from unstructured electronic medical records (EMRs) [7]. Their research involves training 32 prediction models with different combinations of classifiers, tokenization strategies, and attribute selection procedures. The analysis of 85,000 classified sentences taken from EMR data serves as the basis for this work. The best-performing model outperforms rule-based methods with an astounding accuracy of 98.14% and an F-score of 0.981 using Support Vector Machine (SVM) with a combination of unigrams and bigrams as tokens. This research highlights the potential of automated classification systems to accurately assess smoking status from EMRs without manual intervention, thereby enhancing clinical research, healthcare delivery, and public health interventions targeting smoking cessation.

Wang et al. Examine the viability of using magnetic resonance imaging (MRI) with deep learning to distinguish between smoking and non-smoking [8]. 127 participants, including smokers and non-smokers, provided head MRI 3D-T1WI images for the study since smoking is a serious public health concern. A convolutional neural network coupled with a recurrent neural network with long short-term memory architecture (ConvLSTM) and a deep 3D convolutional neural network (Conv3D) were the two deep learning models created. The Conv3D model achieved a promising accuracy of 80.6%, while the ConvLSTM model outperformed, attaining an accuracy of 93.5%. These findings point to the possibility of deep learning-based MRI to outperform conventional support vector machine (SVM) techniques in terms of smoking status prediction accuracy. This strategy has the potential to improve our knowledge of the structural alterations in the brain caused by smoking and to improve smoking cessation programs, which will lessen the impact of smoking-related illnesses on public health.

Haque et al. Determine the validity of smoking status ascertainment methods in population-based electronic health databases by conducting a comprehensive review and meta-analysis [9]. With 57 articles and 116 algorithms analyzed, predominantly based on electronic medical record (EMR) data, The study evaluates these algorithms' efficacy, a number of them make use of diagnosis codes for illnesses connected to smoking. Approximately 50% of the algorithms incorporate machine-learning models. Even though algorithm performance varied greatly, the pooled estimates show a positive predictive value of 0.843, a sensitivity of 0.672, and a specificity of 0.918. Meta-regression analysis identifies model-based algorithms and EMR data as factors associated with higher sensitivity. Some algorithms have limits in terms of sensitivity and positive predictive value; however, those that use several linked data sources and machine learning models show enhanced validity. The present analysis highlights the significance of strong algorithms in accurately determining smoking status in electronic health data, which is crucial for comprehending the dangers associated with smoking and guiding public health measures.

In their study, Vaghefi et al. explore the use of Convolutional Neural Networks (CNNs) to detect smoking status from retinal images, highlighting the impact of smoking on retinal vasculature and its association with cardiovascular disorders [10]. Utilizing a dataset of 165,104 retinal images labelled with self-reported smoking status, the researchers developed CNN models using two pre-processing methods: ”contrast-enhanced” and ”skeletonized”. While the contrast-enhanced model demonstrated an accuracy of 88.88% and a specificity of 93.87%, the skeletonized model showed lower performance. Attention maps generated by the contrast-enhanced model emphasized predictive features such as retinal vasculature, perivascular regions, and the fovea. Despite achieving high accuracy, the study suggests further research to improve sensitivity, potentially by considering smoking frequency, duration, and dosage. This research underscores the potential of CNNs in detecting smoking status from retinal images, providing valuable insights into the physiological changes associated with smoking.

In their study, Dr. N. Palanivel et al. explore the use of Deep Neural Networks (DNN) for the early detection of Alzheimer’s Disease (AD) using a large MRI dataset comprising normal and diseased subjects [11]. AD is a progressive neurodegenerative disorder characterized by cognitive decline and memory loss, making early diagnosis crucial for effective management and intervention. The DNN algorithms developed in this research achieved a high accuracy of approximately 95.1% in predicting AD, surpassing the performance of CNN models. This study addresses the significant clinical, social, and economic need for early detection of AD, offering promising implications for improving patient outcomes and healthcare management.

In their study, Dr. N. Palanivel et al. introduce novel techniques for diabetes classification and prediction in healthcare applications, leveraging deep learning and advanced genetic optimization [12]. The vast amount of data in medical databases poses a challenge for accurate data classification, particularly in disease diagnosis. The study proposes three methods to accurately classify patient medical data by establishing optimal association rules: Advanced Genetic Optimization (AGO)-based deep neural connectivity classification (ABGO-DNAC), integrated gradient-based classification, and amplified classification. These approaches aim to enhance diabetes diagnosis outcomes by improving classification accuracy and reducing processing time. The AGO method specifically focuses on enhancing diabetes classification accuracy by utilizing an expanded evolutionary algorithm to optimize association rules and employing an amplified gradient derivative classifier. Attributes are organized based on IDGBC selection trees, further optimizing the classification process. This research contributes to the advancement of healthcare analytics, offering promising implications for improving diabetes diagnosis and patient care.

3. PROPOSED WORK

A. *Overview*

The proposed work introduces biosignal data for smoking status prediction through the fusion of standard autoencoders and artificial neural networks. On the other hand, no relevant research has been done on the use of biosignal data to forecast smoking status or smoking cessation. A number of crucial elements are included in the suggested system for predicting smoker status using biosignals, including data collection, preprocessing, model construction, and application integration. The various machine learning models have been explored, but comparatively, the model developed using a fusion of an autoencoder and an artificial neural network (ANN) gives good performance. Overall, the proposed system aims to provide a comprehensive and accessible tool for predicting smoker status and promoting smoking cessation using biosignal data and deep learning techniques. Table I. is dataset description.

B. *Dataset Description*

1) *Demographic data*

This category includes Age, Height, Weight, and Waist circumference.

2) *Clinical data*

- Vision: Eyesight measurements (left and right)
- Hearing: Hearing measurements (left and right)
- Cardiovascular: Blood pressure (systolic and relaxation)
- Metabolic: Fasting blood sugar levels, Cholesterol levels, Triglyceride levels, Haemoglobin levels
- Renal: Urine protein levels, Serum creatinine levels
- Liver function: AST (Aspartate Aminotransferase) levels, ALT (Alanine Aminotransferase) levels, and GTP (Gamma-Glutamyl Transferase) levels
- Oral health: Dental caries status

C. *Proposed Algorithm*

The proposed algorithm aims to first extract meaningful features from the input data using an autoencoder and then utilize these features to train a neural network classifier for predicting smoker status. The use of an autoencoder helps in capturing the underlying patterns in the data, potentially

TABLE I. Dataset Description

Feature Category	Feature Name	Description
Demographic Information	Age (years)	User's age in years.
Anthropometric Measurements	Height (cm)	User's height in centimeters.
	Weight (kg)	User's weight in kilograms.
	Waist (cm)	User's waist circumference in centimeters.
Visual acuity	Eyesight (left/right)	It represents the eyesight health
Auditory Perception	Hearing (left/right)	It represents the hearing health
Vital Signs	Systolic Blood Pressure (mmHg)	Systolic blood pressure reading in millimeters of mercury.
	Relaxation Blood Pressure (mmHg)	Diastolic blood pressure reading in millimeters of mercury.
Biochemical Measures	Fasting Blood Sugar (mg/dL)	Blood sugar level measured after an overnight fast (mg/dL).
	Cholesterol (mg/dL)	Total cholesterol level in the blood (mg/dL).
	Triglyceride (mg/dL)	Level of triglycerides (a type of fat) in the blood (mg/dL).
	HDL (mg/dL)	Level of high-density lipoprotein (HDL) or ”good” cholesterol in the blood (mg/dL).
	LDL (mg/dL)	Level of low-density lipoprotein (LDL) or ”bad” cholesterol in the blood (mg/dL).
	Haemoglobin (g/dL)	Haemoglobin level in the blood (g/dL), which carries oxygen.
Renal	Urine protein (mg/dL)	Amount of protein found in the urine (mg/dL).
	Serum creatinine (mg/dL)	Level of creatinine (a waste product filtered by the kidneys) in the blood (mg/dL).
Liver function	AST (U/L)	Aspartate aminotransferase, an enzyme found in the liver, muscles, and red blood cells
	ALT (U/L)	Alanine aminotransferase, another enzyme found in the liver.
	GTP (U/L)	Gamma-glutamyl transferase, an enzyme found in the liver, bile ducts, and pancreas.
Target Variable	Smoker Status (categorical)	Binary value indicating smoker (1) or non-smoker (0).

improving the classification performance of the neural network classifier. The proposed algorithm construction is divided into two phases, namely the autoencoder architecture and building the neural network classifier. Figure 1 shows Proposed Algorithm Architecture.

1) Autoencoder Architecture:

- Defined an autoencoder with an input layer, two hidden layers for encoding (128 neurons, 64 neurons), and a mirrored architecture for decoding.
- Compile the autoencoder using the ’adam’ optimizer and ’mean-squared-error’ loss function.
- Fit the autoencoder to the standardized training data for 50 epochs with a batch size of 32.
- Validate the autoencoder on the standardized test data.
- Use the encoder part of the trained autoencoder to extract features from the standardized training and test data.

2) Building the Neural Network Classifier:

- Created a sequential model for classification.
- Add a dense layer with 128 neurons and ’relu’ activation function, a dropout layer with 0.5 dropout rate, another dense layer with 64 neurons and ’relu’ activation function, and another dropout layer.
- Add a dense output layer with 1 neuron and ’sigmoid’ activation function for binary classification.
- Compile the model using the ’adam’ optimizer, the ’binary-crossentropy’ loss function, and the accuracy metric.
- Validate the encoded test features by training the model with the encoded features for 50 epochs at a batch size of 32.

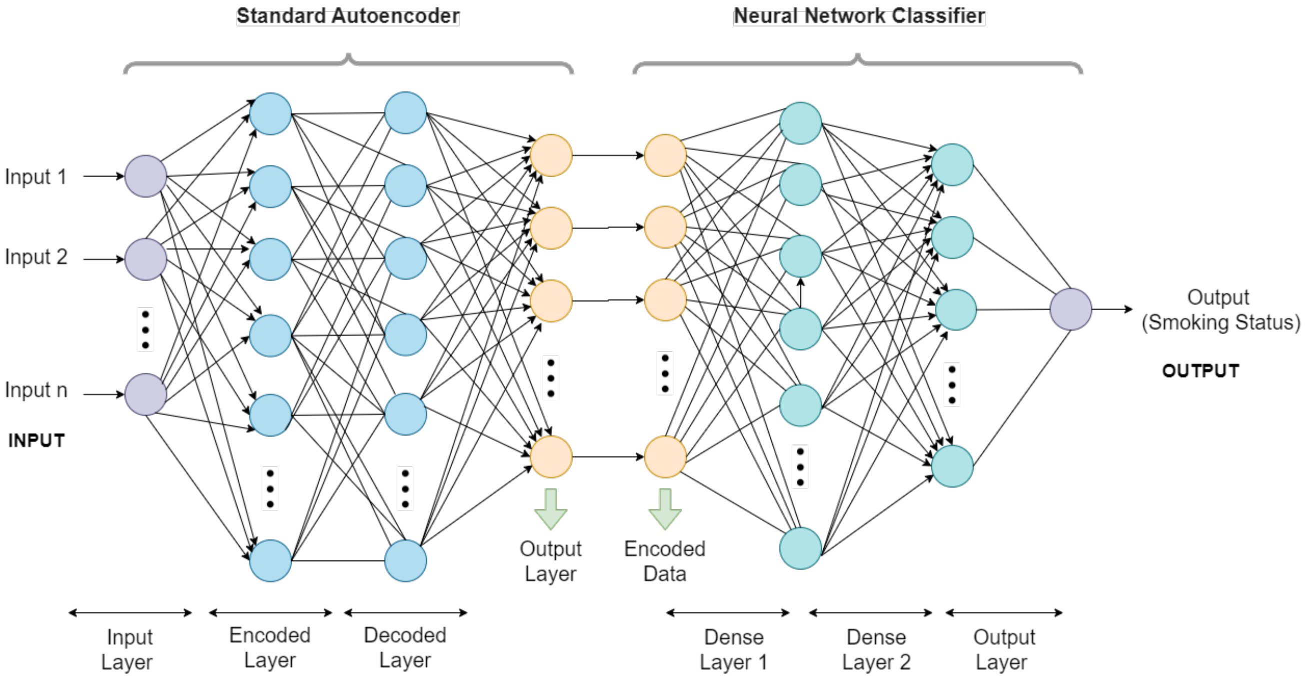


Figure 1. Proposed Algorithm Architecture

- Finally, assessing the accuracy and loss measures of the model's performance on the test set.

D. *Workflow of the Algorithm*

1) *Data Acquisition and Preparation:*

Obtaining bio-signal data for smokers and non-smokers. This data includes demographic data and clinical data. Ensure the data is properly labeled, indicating whether each data point belongs to a smoker or a non-smoker. Clean and preprocess the data, which involves handling missing values, outliers, or inconsistencies.

2) *Data Splitting:*

Dividing labeled data into two sets: training and testing data. The training data (typically 70–80% of the total data) will be used to train the model. The testing data (remaining 20–30%) will be used to evaluate the model's performance on unseen data.

3) *Autoencoder Training:*

Build and train the autoencoder model using the scaled training data. Train the autoencoder to reconstruct the original data from the encoded features as accurately as possible. This process helps the autoencoder learn underlying patterns and relationships within the biosignals.

4) *Feature Extraction (Using Encoder):*

Once trained, use the encoder part of the autoencoder to extract the encoded features from both the scaled training data and the original testing data. These encoded features represent compressed versions of the original bio-signals, hopefully capturing the most important information for smoker classification.

5) *Neural Network Classifier Training:*

Build a neural network classifier model and train it on the extracted encoded features from the training data and the corresponding labels (smoker or non-smoker). The neural network classifier learns to associate specific patterns in the encoded features with the smoker and non-smoker labels.

6) *Model Evaluation:*

Evaluate the performance of the trained neural network classifier on the encoded features from the testing data. Common metrics used for classification include accuracy, precision, recall, F1-score and AUC-ROC score.

4. METHODOLOGY

A. *Data Source*

The overview that outlines the specific measurements obtained for each biosignal data, including demographic information, vision and hearing assessments, physiological parameters, lipid panel and complete blood count tests, urine protein levels, serum creatinine levels, liver function tests, and dental caries status. Each of the biosignal data plays a crucial role in informing the predictive model, contributing essential insights into the individual's health status and smoking behaviour. Table II discusses the dataset source and their tests for collecting the biosignal data.

B. *Data Preprocessing*

Data preprocessing encompasses various techniques that clean, transform, and format the data to make it suitable

TABLE II. Data Source

Data Source	Description
Demographic Data	Age, height, weight, and waist measurements are typically obtained through direct measurement or self-reporting.
Vision Measurements	Eyesight measurements, often in the form of a ”Snellen fraction,” are assessed using eye charts like the Snellen Eye Chart or the Random E Chart. Results are expressed as a fraction (e.g., 20/20).
Hearing Measurements	Hearing assessments may involve tests like pure-tone audiometry or tympanometry, administered by an audiologist or health professional.
Blood Pressure	Blood pressure readings, including systolic and diastolic measurements, are obtained using a sphygmomanometer.
Blood Sugar Levels	A glucometer is a medical gadget that is specifically used to measure blood sugar levels during fasting.
Lipid Panel	Lipid panels are blood tests used to measure triglyceride and cholesterol levels.
Complete Blood Count	A complete blood count (CBC) test is conducted to assess haemoglobin levels, a protein in red blood cells. Low levels may indicate anaemia.
Urine Protein Level	Urine protein levels are qualitatively assessed using a urinalysis dipstick, providing results ranging from negative (normal) to 4+ (very high protein levels detected).
Serum Creatinine	Serum creatinine levels are measured through a blood test, typically drawn from the arm, to evaluate kidney function.
Liver Function Test	Liver function tests measure values of AST, ALT, and GTP levels to assess liver health and functioning.
Dental Caries Status	Dental caries status is evaluated through visual-tactile examination by a dentist or dental hygienist

for model training. One essential step in this process is Normalization using techniques like StandardScaler.

Normalization with StandardScaler:

StandardScaler transforms each feature in the data (bio-signals) to have a mean of 0 and a standard deviation of 1. Prior to feeding the bio-signal characteristics into the autoencoder and neural network classifier, StandardScaler is utilized to normalize them. As a result, the model's ability to distinguish between smokers and non-smokers may be enhanced and all bio-signals are guaranteed to contribute equally to its learning process. By bringing all features down to a common scale, this effectively prevents any one feature from controlling the learning process by having a significantly bigger or smaller scale than the rest. StandardScaler divides the output by the feature's standard deviation after subtracting each feature's mean from each data point. Here, x_scaled represents the normalized value of the feature x.

In terms of math, for a feature x:

$$x\_scaled = \frac{x - mean(x)}{standard\_deviation(x)}$$

There are several advantages to using StandardScaler before feeding data into a machine learning model:

1) Improved Model Performance:

Numerous machine learning methods may be sensitive to the size of the features, particularly those that compute the distances between data points (such as certain classification algorithms). Normalization guarantees that every feature makes an equal contribution to the model's computations, which may improve convergence and performance.

2) Faster Training:

When features have different scales, the model might take longer to learn because it needs to adjust for these differences. Normalization can help the model converge faster as the features are presented on a similar scale.

3) Prevents Feature Domination:

If a feature has a much larger scale compared to others, it can overshadow the influence of the other features on the model's learning process. Normalization prevents this by putting all features on a level playing field.

C. Deep Learning Architecture

1) Autoencoder for Feature Extraction and Dimensionality Reduction

Standard autoencoders use Mean Square Error for reconstruction, which offers a general-purpose approach for dimensionality reduction and feature extraction. Standard autoencoder that doesn't specifically target any particular distributions or noise management techniques, instead learning a compressed representation of the biosignal data. This process not only helps in reducing the computational complexity of the subsequent modelling steps but also ensures that the most relevant features are retained for accurate prediction. Figure 2 shows Standard Autoencoder structural diagram.

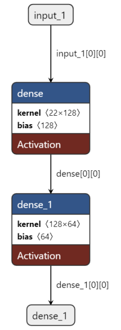


Figure 2. Standard Autoencoder structural diagram

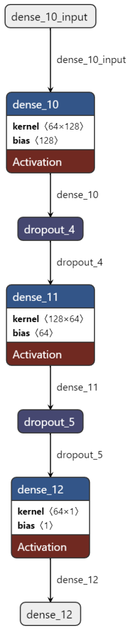


Figure 3. Artificial Neural Network structural diagram

2) *ANN for Prediction and Classification:*

ANN is a broad term encompassing various neural network architectures, the network proposed is Dense Feed-forward Neural Network. It accepts the encoded data representing bio-signals after processing by the autoencoder. Composed of densely connected layers with ReLU activation functions. These layers learn complex patterns within the bio-signal data that may correlate with smoking status. A single neuron that functions as a sigmoid activator. By converting the output from the hidden layers into a value between 0 and 1, it determines the likelihood that the individual smokes. Dense feedforward neural networks are well-suited for classification tasks on structured data, like the encoded biosignal features. Figure 3 shows Artificial Neural Network structural diagram.

D. Flask Application Development

In the realm of smoker status prediction and smoking cessation initiatives, the development of a Flask application serves as a pivotal tool for delivering actionable insights and facilitating user engagement. By leveraging the versatility and scalability of Flask, Intuitive and user-friendly platform that has been created which empowers individuals to assess their smoker status, access resources for smoking cessation, and gain valuable insights into the benefits of quitting smoking.

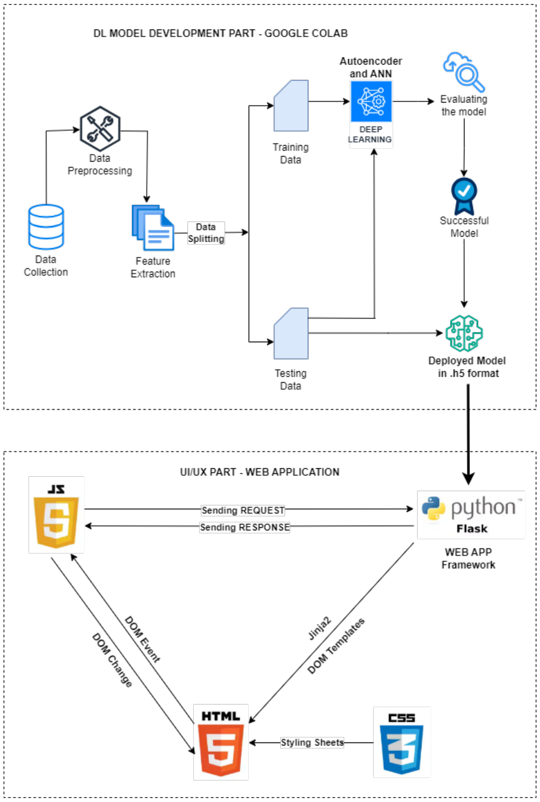


Figure 4. System Architecture Diagram

E. System Architecture

The system architecture is a Flask web application for smoker status prediction with chatbot integration. The user interacts with the web application through a web browser, likely filling out a form with biosignal data. A Flask backend receives the data, utilizes a deep learning model (autoencoder and ANN) for prediction, and delivers the results along with the prediction value and various web pages for smoking cessation awareness through the web browser. Figure 4 is the pictorial representation of System Architecture.

5. RESULTS

Using the biosignal dataset, the effectiveness of several machine learning (ML) methods and the deep learning (DL) model for predicting smoker status was assessed. The outcomes show the advantages and disadvantages of each strategy, shedding light on whether DL is a better option than conventional ML methods.

TABLE III. Performance of various ML Algorithm

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-score	AUC-ROC score
Logistic Regression	74	74	71	71	71	74
Gaussian Naive Bayes	72	72	66	78	71	73
Decision Tree	100	70	66	65	65	69
K Nearest Neighbors	77	74	70	73	71	74
Random Forest Classifier	100	77	72	80	76	78
Adaboost Classifier	77	77	72	78	75	77
Gradient Boosting Classifier	78	78	72	82	77	78

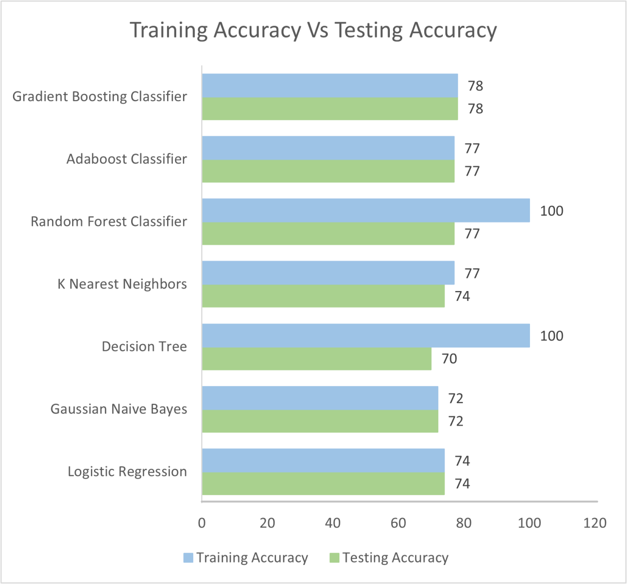
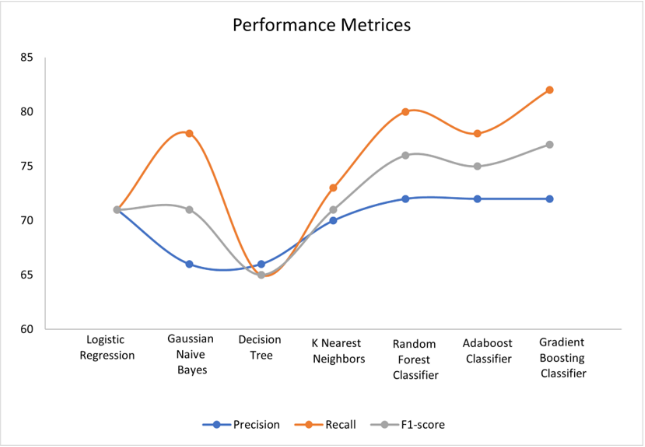


Figure 5. Training Accuracy Vs Testing Accuracy



$$Precision = \frac{TP}{(TP+FP)} \qquad Recall = \frac{TP}{(TP+FN)} \qquad F1\ Score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)}$$

TP – True Positive FP – False Positive FN – False Negative

Figure 6. Performance metrices of various ML model

A. Machine Learning Algorithm Performance:

Using the biosignal dataset, the effectiveness of several machine learning (ML) algorithms for predicting smoking status was assessed. The outcomes reveal that the Gradient Boosting Classifier attained the highest testing accuracy of 78%, trailed closely by the Adaboost Classifier and Random Forest Classifier, also achieving an accuracy of 77%. Table III discusses the performance of various ML algorithms with evaluation metrics.

The machine learning algorithms were tested for accuracy, with Random Forest achieving the highest score (100%) but potentially overfitting the data, while Decision Tree shows similar performance but with a testing accuracy of 70%. Figure 5 is the graphical representation, which depicts the Training accuracy versus Testing accuracy of various ML algorithms.

In an evaluation of various machine learning algorithms on a biosignal dataset, three algorithms emerged as the top performers. Adaboost, Gradient Boosting, and Random Forest classifiers all achieved a precision score of 72%. When it comes to recall, the Gradient Boosting Classifier takes the crown with a score of 82%, followed by Random Forest at 80%. Combining precision and recall, the Gradient Boosting Classifier again secures the top spot with an F1-score of 77%, with Random Forest trailing closely behind at 76%. Overall, these results suggest that Gradient Boosting and Random Forest are strong contenders for biosignal analysis tasks, with Adaboost also showing promise for accurate classification. Figure 6 is the graphical representation, which depicts the performance metrices such as precision, recall and F1 score of various ML algorithms.

However, AUC-ROC offers a robust and informative metric for evaluating binary classification models, considering different thresholds, and comparing various models. The highest score is for Logistic Regression (78%), while the lowest score is for Decision Tree (69%). It's important to note that the overall performance of all models is relatively low, as none of the AUC-ROC scores are above 80%. Hence, the proposed approach goes on with a deep learning model. Figure 7 is the graphical representation, which

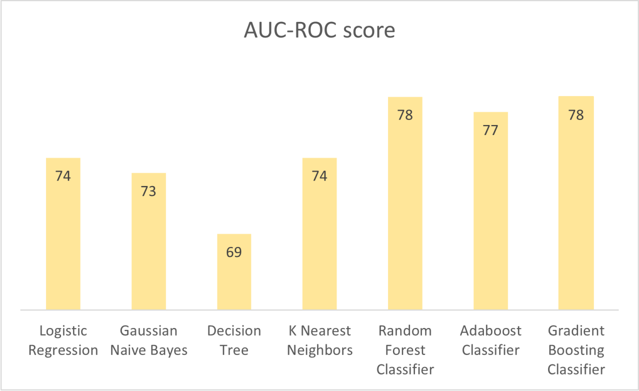


Figure 7. AUC ROC Curve of various ML model

depicts the AUC ROC performance metrices of various ML algorithms.

B. Deep Learning Algorithm Performance:

The evaluation metrics provided indicate that the model is performing reasonably well for predicting a person's smoking status for cessation using biosignals. Here's a brief analysis of each metric:

1) Accuracy:

The accuracy of 0.80 indicates that the model is correctly predicting the smoking status of individuals approximately 80% of the time. While the accuracy score is commendable, it's crucial to contextualize it within the problem domain and consider the implications of false positives and false negatives.

2) Precision:

Precision measures the proportion of positive predictions that are correct. A precision of 0.710 means that when a model predicts a person as a smoker, it is correct about 71% of the time. This indicates that the model has a relatively low rate of false positives.

3) Recall:

Recall, often referred to as sensitivity, gauges the percentage of true smokers that the model correctly identifies. A recall of 0.835 means that the model is able to correctly identify about 83.5% of actual smokers. This indicates that the model has a relatively low rate of false negatives.

4) F1-score:

The F1-score provides a balance between precision and recall by taking the harmonic mean of the two measures. The model appears to have an excellent balance between recall and precision, as indicated by its 77% F1-score.

5) AUC-ROC score:

The AUC-ROC score evaluates the area under the receiver operating characteristic (ROC) curve, reflecting the model's capacity to differentiate between smokers and non-smokers. An AUC-ROC score of 86% suggests that the model demonstrates strong classification performance. Figure 8 is the graphical representation, which depicts the AUC ROC performance of Deep Learning model.

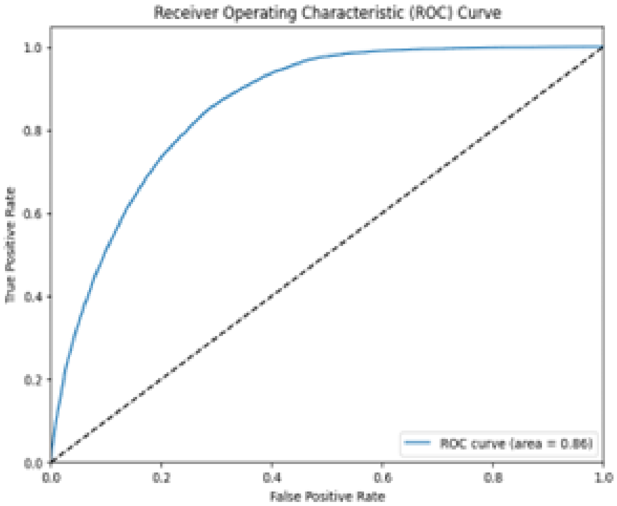


Figure 8. AUC ROC Curve of Deep Learning model

When comparing machine learning methods to the deep learning model, findings indicate that the deep learning model achieved an accuracy rate of 80%, comparable to the 77% accuracy rates of both the Gradient Boosting Classifier and Random Forest Classifier. But in terms of precision, recall, and F1-score, deep learning - a hybrid of conventional autoencoder and artificial neural network (ANN) models - performed better than these approaches.

In contrast to the majority of machine learning algorithms, the deep learning model's precision of 71% shows that it can accurately identify smokers while reducing false positives. The recall of the DL model was 84%, indicating its effectiveness in capturing true positive instances of smokers. The F1-score of the DL model was 77%, which is higher than most ML algorithms, indicating its overall effectiveness in correctly classifying smokers while maintaining a balance between precision and recall.

6. DISCUSSION

A. Practical Implication

Case Study 1: Smoking Cessation Program in a Corporate Environment

Background: A large technology company employs over 5,000 staff members and has observed a high prevalence of smoking among its employees. This has led to increased healthcare costs and reduced productivity. To address these issues, the company decided to implement a comprehensive smoking cessation program leveraging the smoking status prediction model developed in this project.

Implementation: The company integrated the smoking status prediction model into their employee wellness portal.

Employees were encouraged to participate in the program by providing their biosignal data. The model, using the autoencoder and ANN, provided each participant with a personalized smoking status report, highlighting their likelihood of being smokers.

Intervention: Based on the smoking status predictions, employees identified as likely smokers were offered personalized smoking cessation plans. These plans included access to counseling services, nicotine replacement therapies, and regular follow-up sessions. The wellness portal also featured educational content on the benefits of quitting smoking, encouraging employees to take action.

Outcomes:

- After six months of implementing the program, the company reported significant improvements in employee health outcomes.
- A 30% reduction in the number of employees who smoked.
- Decreased healthcare costs related to smoking-related illnesses by 15%.
- Increased overall productivity and reduced absenteeism.

Case Study 2: Community Health Initiative in a Rural Area

Background: In the rural community of Greenfield, smoking rates are significantly higher than the national average. The local health department launched a community health initiative aimed at reducing smoking prevalence and improving public health. They decided to utilize the smoking status prediction model to tailor their interventions.

Implementation: The health department set up mobile health clinics that traveled to different parts of Greenfield. Residents were invited to visit the clinics and undergo health screenings, where their biosignal data were collected. The smoking status prediction model was used to analyze the data and provide individualized smoking risk assessments.

Intervention: Residents identified as likely smokers received personalized intervention plans. These included access to free or subsidized smoking cessation resources, such as nicotine patches, gum, and support groups. The health department also organized workshops and community events to educate residents about the dangers of smoking and the benefits of quitting.

Outcomes:

- Within a year of launching the initiative, Greenfield experienced notable public health improvements.
- Smoking rates in the community decreased by 25
- There was a significant drop in the incidence of smoking-related diseases, such as chronic obstructive pulmonary disease (COPD) and heart disease.
- Increased awareness about the risks of smoking and the benefits of quitting among community members.

Discussion of Practical Implications:

These case studies illustrate the potential real-world applications of the smoking status prediction model developed in this project. By providing personalized risk assessments, the model can help target interventions more effectively, leading to improved health outcomes. In corporate settings, this can translate to lower healthcare costs, higher productivity, and better employee well-being. In community health initiatives, it can result in reduced smoking rates, a lower incidence of smoking-related diseases, and increased public health awareness. By demonstrating the effectiveness of personalized smoking cessation plans based on accurate risk predictions, this project highlights the valuable role of advanced machine learning models in public health interventions. The integration of such models into wellness programs and health initiatives can significantly contribute to achieving better health outcomes and promoting a healthier society.

B. Key features of the web application

1) Smoker Status Prediction:

The core functionality of the Flask application revolves around smoker status prediction based on the provided dataset. Users can input relevant health metrics such as age, weight, cholesterol levels, and more, and receive a personalized prediction of their smoker status. This prediction is powered by the deep learning model integrating autoencoder and ANN architectures, ensuring high accuracy and reliability.

2) Interactive Visualization:

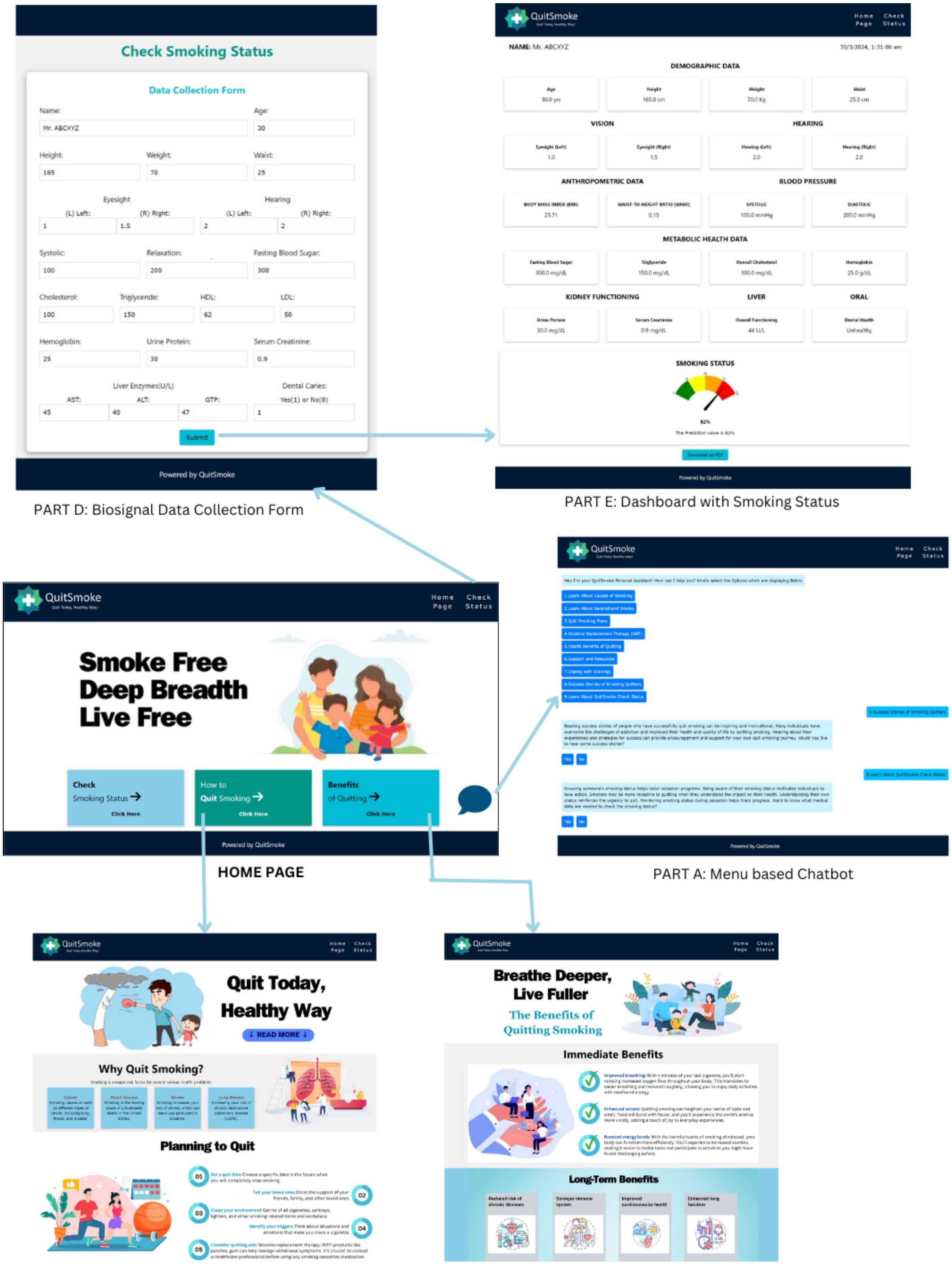
To enhance user experience and facilitate data interpretation, the Flask application incorporates interactive visualization tools. Users can explore their health metrics through dynamic charts, graphs, and visualizations, allowing them to gain deeper insights into the factors influencing their smoker status.

3) Educational Resources:

The Flask application has parts devoted to offering thorough information on the dangers of smoking, the advantages of stopping, and evidence-based cessation techniques because it recognizes the critical role that education plays in smoking cessation efforts.

4) Why Quit Smoking Page:

A dedicated page within the application is designed to educate users about the compelling reasons to quit smoking. Through compelling narratives, testimonials, and statistics, this page reinforces the importance of smoking cessation and motivates users to take proactive steps towards quitting.



Check Smoking Status

Data Collection Form

Name: Mr. ABCXYZ			Age: 30
Height: 165	Weight: 70		Waist: 25
Eyesight (L) Left: 1	Eyesight (R) Right: 1.5	Hearing (L) Left: 2	Hearing (R) Right: 2
Systolic: 100	Relaxation: 200		Fasting Blood Sugar: 300
Cholesterol: 100	Triglyceride: 150	HDL: 62	LDL: 50
Hemoglobin: 25	Urine Protein: 30		Serum Creatinine: 0.9
Liver Enzymes(U/L) AST: 45	ALT: 40	GTP: 47	Dental Caries: Yes(1) or No(0): 1

PART D: Biosignal Data Collection Form

NAME: Mr. ABCXYZ 10/3/2024, 1:31:06 am

DEMOGRAPHIC DATA

Age	Height	Weight	Waist
30.0 yrs	165.0 cm	70.0 Kg	25.0 cm

VISION / HEARING

Eyesight (Left)	Eyesight (Right)	Hearing (Left)	Hearing (Right)
1.0	1.5	2.0	2.0

ANTHROPOMETRIC DATA / BLOOD PRESSURE

BODY MASS INDEX (BMI)	WAIST-TO-HEIGHT RATIO (WHtR)	SYSTOLIC	DIASTOLIC
25.71	0.15	100.0 mmHg	200.0 mmHg

METABOLIC HEALTH DATA

Fasting Blood Sugar	Triglyceride	Overall Cholesterol	Hemoglobin
300.0 mg/dL	150.0 mg/dL	100.0 mg/dL	25.0 g/dL

KIDNEY FUNCTIONING / LIVER / ORAL

Urine Protein	Serum Creatinine	Overall Functioning	Dental Health
30.0 mg/dL	0.9 mg/dL	44 U/L	Unhealthy

SMOKING STATUS

82%

The Prediction value is 82%

PART E: Dashboard with Smoking Status

Smoke Free
Deep Breadth
Live Free

Check Smoking Status →

How to Quit Smoking →

Benefits of Quitting →

HOME PAGE

Hey I'm your QuitSmoke Personal Assistant!! How can I help you? Kindly select the Options which are displaying Below

1.Learn About Causes of Smoking
2.Learn About Secondhand Smoke
3.Quit Smoking Plans
4.Nicotine Replacement Therapy (NRT)
5.Health Benefits of Quitting
6.Support and Resources
7.Coping with Cravings
8.Success Stories of Smoking Quitters
9.Learn About QuitSmoke Check Status

8.Success Stories of Smoking Quitters

Reading success stories of people who have successfully quit smoking can be inspiring and motivational. Many individuals have overcome the challenges of addiction and improved their health and quality of life by quitting smoking. Hearing about their experiences and strategies for success can provide encouragement and support for your own quit-smoking journey. Would you like to hear some success stories?

9.Learn About QuitSmoke Check Status

Knowing someone's smoking status helps tailor cessation programs. Being aware of their smoking status motivates individuals to take action. Smokers may be more receptive to quitting when they understand the impact on their health. Understanding their own status reinforces the urgency to quit. Monitoring smoking status during cessation helps track progress. Want to know what medical data are needed to check the smoking status?

PART A: Menu based Chatbot

Quit Today,
Healthy Way

Why Quit Smoking?

Smoking is a major risk factor for several serious health problems

Cancer	Heart disease	Stroke	Lung disease
Smoking causes at least 13 different types of cancer, including lung, throat, and bladder.	Smoking is the leading cause of preventable death in the United States.	Smoking increases your risk of stroke, which can leave you paralyzed or disabled.	Increasing your risk of chronic obstructive pulmonary disease (COPD).

Planning to Quit

01 Set a quit date Choose a specific date in the future when you will completely stop smoking.

02 Tell your loved ones Enlist the support of your friends, family, and other loved ones.

03 Clean your environment Get rid of all cigarettes, ashtrays, lighters, and other smoking-related items and workplace.

04 Identify your triggers Think about situations and emotions that make you crave a cigarette.

05 Consider quitting aids Nicotine replacement therapy (NRT) products like patches, gum can help manage withdrawal symptoms. *It's crucial to consult a healthcare professional before using any smoking cessation medication.*

PART B: How to Quit Smoking Page UI

Breathe Deeper,
Live Fuller

The Benefits of Quitting Smoking

Immediate Benefits

✓ Improved breathing: Within minutes of your last cigarette, you'll start noticing increased oxygen flow throughout your body. This translates to easier breathing and reduced coughing, allowing you to enjoy daily activities with newfound energy.

✓ Enhanced senses: Quitting smoking can heighten your sense of taste and smell. Food will burst with flavor, and you'll experience the world's aromas more vividly, adding a touch of joy to everyday experiences.

✓ Boosted energy levels: With the harmful toxins of smoking eliminated, your body can function more efficiently. You'll experience increased stamina, making it easier to tackle tasks and participate in activities you might have found challenging before.

Long-Term Benefits

Reduced risk of chronic diseases	Stronger immune system	Improved cardiovascular health	Enhanced lung function

PART C: Benefits of Quitting Page UI

Figure 9. Web Application UI Structure

5) Benefits of Quitting Smoking Page:

Another integral component of the Flask application is the Benefits of Quitting Smoking page, which outlines the myriad health benefits associated with smoking cessation. Users can research how quitting smoking can enhance their overall health, reduce their chance of getting chronic illnesses, and increase their quality of life.

6) Menu-Based Chatbot:

To address user queries and provide personalized support, the Flask application integrates a menu-based chatbot. Users can interact with the chatbot to seek answers to common questions about smoking cessation, receive tips for quitting, and access additional resources for support.

C. Benefits of the Web App

1) Accessibility:

The web application makes smoker status prediction readily accessible to a wider audience. Users can conveniently access the application from any web browser without requiring specialized software installation.

2) Ease of Use:

The user interface is designed to be intuitive and user-friendly, allowing individuals with varying levels of technical expertise to easily interact with the model and obtain predictions.

3) Personalized Feedback:

The application provides personalized predictions based on the user's unique Biosignal data. This can be more impactful than generic information about smoking risks.

4) Educational Support:

By integrating educational content based on the predicted smoker status, the web app can potentially serve as a valuable tool for raising awareness about smoking risks and promoting smoking cessation efforts.

7. CONCLUSION

In conclusion, our research endeavours to revolutionize smoking status prediction and smoking cessation interventions by pioneering the integration of biosignal data and deep learning techniques. By leveraging a vast dataset of 1.5 lakhs biosignal data points, we have developed a robust predictive model that surpasses existing methodologies in accuracy and performance. Through the fusion of standard autoencoders and artificial neural networks, our model demonstrates its effectiveness in precisely determining smoker status with an amazing 80% accuracy and 86% AUC-ROC score. The deep learning models (autoencoder and ANN) outperformed traditional ML algorithms. Their higher recall and accuracy were a result of their capacity to recognize subtle patterns, process high-dimensional data, and understand complicated features.

Moreover, our comprehensive approach extends beyond model development, encompassing crucial stages such as data collection, preprocessing, model construction, and application integration. By addressing these fundamental elements, we ensure the reliability, scalability, and practicality of our proposed system. Furthermore, our user-friendly Flask frontend, featuring dedicated pages for Smoker Status Prediction, Why Quit Smoking, and Benefits of Quitting Smoking, functions as a valuable asset for individuals aiming to comprehend and address their smoking habits. Additionally, the integration of a menu-based chatbot enhances user engagement and accessibility, providing personalized support and information tailored to each user's needs. By empowering individuals with insights into the detrimental effects of smoking and the benefits of cessation, our system aims to catalyse positive behaviour change and promote healthier lifestyles.

References

[1] Y. K. J.-F. Etter, G. Vera Cruz, "Predicting smoking cessation, reduction and relapse six months after using the stop-tabac app for smartphones: a machine learning analysis," *BMC Public Health*, vol. 23, p. 1, 2023.

[2] L. M. L. T. T. T. L. A. C. T. J. L. Y. M. R. M. D. L. D. T. Issabakhsh, Mona; Sánchez-Romero, "Machine learning application for predicting smoking cessation among us adults: An analysis of waves 1-3 of the path study," *PLoS One*, vol. 18, p. 6, 2023.

[3] D. Nuryunarsih *et al.*, "Artificial neural network machine learning prediction of the smoking behavior and health risks perception of indonesian health professionals," *Environmental Analysis Health and Toxicology*, vol. 38, no. 1, pp. Article e2 023 003–0, 2023.

[4] C. Frank *et al.*, "Predicting smoking status using machine learning algorithms and statistical analysis," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 2, pp. 184–189, 2018.

[5] S. S. Thakur *et al.*, "Real-time prediction of smoking activity using machine learning based multi-class classification model," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 14 529–14 551, 2022.

[6] C.-C. Lai *et al.*, "Development of machine learning models for prediction of smoking cessation outcome," *International Journal of Environmental Research and Public Health*, vol. 18, no. 5, p. 2584, 2021.

[7] A. Caccamisi *et al.*, "Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records," *Scandinavian Journal of Public Health*, vol. 48, no. 3, pp. 316–324, 2020.

[8] S. Wang *et al.*, "Discrimination of smoking status by mri based on deep learning method," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 11, pp. 1113–1120, 2018.

[9] M. A. Haque *et al.*, "The validity of electronic health data for measuring smoking status: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 24, p. Article no. 33, 2024.

[10] E. Vaghefi *et al.*, "Detection of smoking status from retinal images: a convolutional neural network study," *Scientific Reports*, vol. 9, p. Article no. 7180, 2019.

[11] R. Indumathi, N. Palanivel, V. Kumar, S. Kannan, and J. M. Ahmed, “Alzheimer’s disease detection using deep neural network,” *TELEMATIQUE*, vol. 21, no. 1, 2022.

[12] V. Selvi, G. Saraswathi, R. Indumathi, N. Palanivel, S. Reshma, and R. Subiksha, “Deep learning based diabetes classification and prediction for healthcare applications using advanced genetic optimization,” *International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6, 2023.