

Explainable Random Forest Framework for Multi-Class Lifestyle Disease Risk Prediction Using BRFSS Data

Prema P.¹, Nivetha P.², Nagasakthi N.³

¹ Assistant Professor, ^{2,3} Student, Department of Computer Science and Engineering,
NPR College of Engineering and Technology, Dindigul, India.

E-mail: ¹premap@nprcolleges.org, ²nagasakthin920822104302@nprcolleges.org, ³nivethap920822104068@nprcolleges.org

Abstract

With the increase in lifestyle diseases globally, there is a need for an effective machine learning approach to predict people who are at risk. In this research, a machine learning framework for multiclass prediction of risk of lifestyle disease using the BRFSS 2024 dataset is developed. There are six risk groups: Healthy, Diabetes Risk, Hypertension Risk, Obesity Risk, Heart Disease Risk, and Kidney Disease Risk. Data preprocessing in the proposed framework is done by treating missing values, converting categorical variables into numeric variables, eliminating duplicate and inconsistent entries, normalizing using the Z-score method, and correlation-based feature selection. Synthetic Minority Over-sampling Technique (SMOTE) is implemented to solve the problem of class imbalance. There are five classifiers tested using the 80/20 ratio for training and testing. They include Logistic Regression, SVM, KNN, Decision Tree, and Random Forest. The analysis reveals that the Random Forest performs well with an accuracy rate of 94.8%, and thus, is selected as the final model for predicting the diseases. Age, smoking, and sleep time have also been identified as important features using feature importance analysis that play an important role in predicting the risk of developing lifestyle diseases..

Keywords: Explainable Machine Learning, Lifestyle Disease Risk Prediction, Health Behaviour Analysis, Ensemble Classification, Synthetic Data Balancing, Feature Importance Analysis, Preventive Recommendation Generation.

1. Introduction

Lifestyle diseases have become a significant health issue because of its connection with demographic, physiologic, and behavioral factors. The factors that could include age, exercise, sleeping hours, smoking, alcohol consumption, diet behavior, and many other health-related attributes altogether might affect the disease risk profile of a person. The availability of such health survey data has enabled the use of computer-based screening methods that can utilize this kind of heterogeneous health data to help predict the risks of developing certain diseases. Machine learning is a very suitable choice for such an environment because many health factors can be considered simultaneously.

Constructing a robust lifestyle disease prediction model needs a proper data processing and modelling pipeline. The health survey data might be characterized by missing values, categorical features, noisy data entries, redundancy, and skewed classes. All these properties might impact the learning phase and thereby impact the predictive power of the model. Thus, in the proposed model architecture, we have included the following preprocessing steps before building the models: data cleaning, median imputation, categorical encoding, Z-score normalization, feature selection and class balancing. Several supervised machine learning algorithms such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest are considered in this experiment setup.

The proposed framework intends to enhance the prediction by considering the elements of model interpretability and preventive recommendation generation. Random Forest is incorporated into the framework as the main prediction model, whereas feature importance is applied in order to determine the health and lifestyle characteristics associated with the prediction results. The generated risk profile is linked with a recommendation engine which generates recommendations regarding diet, exercise, stress management, and visiting doctor. The entire process is realized within one web-based framework.

2. Literature Review

Machine learning techniques have been employed in predicting and assessing health conditions due to their capability to establish model relationships between different health parameters. There has been an increase in the number of noncommunicable diseases and

cardiovascular conditions in the world, prompting the development of computing techniques to predict risks and provide preventive healthcare [1], [2], [5]. Several comparative studies have been carried out to assess the applicability of traditional supervised learning models to predict diseases, which include Logistic Regression, SVM, Decision Tree, KNN, and ensembles [4], [8], [9].

Multiple disease prediction and ensemble methods for cardiovascular risk prediction have also been studied in recent research, which demonstrates the potential for unified machine learning approaches to predict various diseases [9], [10]. Efficient gradient boosting-based predictions can be done using algorithms like LightGBM, and systematic model parameter tuning can be performed using optimization frameworks like Optuna [3], [6].

Machine learning has also been studied for its application in security-related fields like deep neural network attacks, demonstrating another area of intelligent learning in the context of dependable computing systems [7]. Nonetheless, the above-mentioned studies are more focused on individual disease prediction, classification, or individual components of the methodology. The current research combines lifestyle health parameters, data preprocessing, class imbalance, comparison of machine learning algorithms, Risk Prediction using Random Forest Algorithm, and preventive measures into one integrated web-based system.

3. Proposed Methodology

3.1 System Architecture and Operational Workflow

The proposed disease risk prediction system architecture is illustrated in Figure 1. The architecture comprises five different layers, namely user, logic, intelligence, persistence, and output/recommendation layers. The user layer serves as the interface to provide inputs on demographics, health, and lifestyle details. Input validation is done prior to sending the input attributes to the application layer.

The logic layer utilizes Flask and takes care of user authentication and session management as well as communication between interface and the prediction service. The intelligence layer forms an interface between the application and the machine learning model that is trained. In the process of prediction, the validated user attributes are subjected to the required preprocessing step and fed into the trained Random Forest algorithm.

The persistence layer includes MySQL for user details and MongoDB for prediction entries and historical health reports. The output layer receives the risk class prediction and recommends preventive measures. The following six prediction classes are taken into consideration in the system architecture: Healthy, Diabetes Risk, Hypertension Risk, Obesity Risk, Heart Disease Risk, and Kidney Disease Risk.

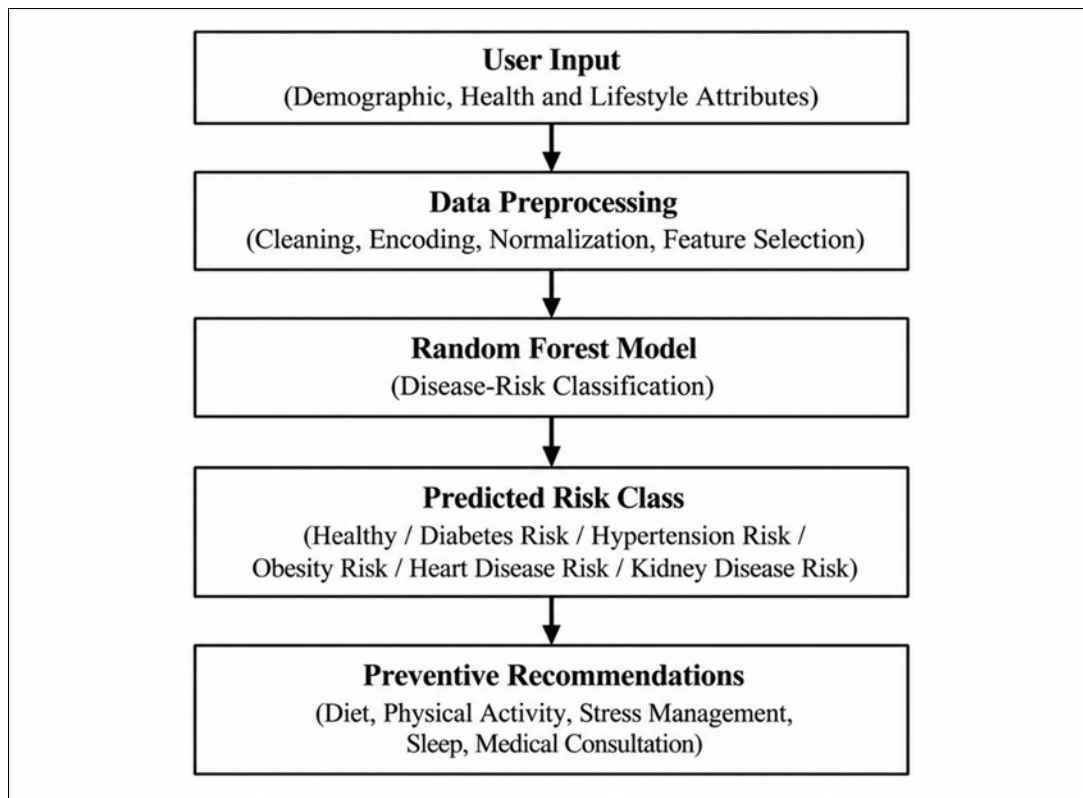


Figure 1. Integrated architecture of the proposed lifestyle disease risk prediction system

3.2 Dataset Description

The proposed framework uses the Behavioral Risk Factor Surveillance System (BRFSS) as the primary source of data [11]. BRFSS is a large-scale health survey, which includes data relevant to health status, risk behaviors, preventive measures, demographics, exercise, smoking, alcohol, and other health-related features. The data set includes heterogeneous features pertaining to demographics, anthropometrics, health and behavior, which are used to create the input features for predicting the risk of developing a lifestyle disease.

Prediction is framed as a multi-class classification of risk associated with lifestyle disease. The input variables include demographic and health characteristics like age, gender, height,

weight, body mass index, blood pressure, blood sugar level, physical activity, hours of sleep, smoking, alcohol intake, dietary habits, stress level, and medical history of family. The output of the prediction includes six pre-defined classes as Healthy, Diabetes Risk, Hypertension Risk, Obesity Risk, Heart Disease Risk, and Kidney Disease Risk. The above classes constitute the target for training and evaluation of five machine learning classifiers. The complete set of characteristics, input variables, experiment setting, and output classes are outlined in Table 1.

Table 1. BRFSS dataset characteristics and experimental configuration

Characteristic	Description
Dataset	Behavioral Risk Factor Surveillance System (BRFSS)
Survey year	2024
Total samples	457,670
Original variables	345
Training–testing ratio	80:20
Training samples	366,136
Testing samples	91,534
Data collection	Combined landline and cellular telephone survey
Demographic attributes	Age, gender
Anthropometric attributes	Height, weight, BMI
Health attributes	Blood pressure, blood sugar, family medical history
Lifestyle attributes	Physical activity, sleep, smoking, alcohol consumption, dietary behaviour, stress level
Prediction task	Multi-class lifestyle disease-risk classification
Prediction labels	Healthy, Diabetes Risk, Hypertension Risk, Obesity Risk, Heart Disease Risk, Kidney Disease Risk

The selected attributes provide a collective representation of demographic, physiological, and behavioral aspects for disease-risk modeling. The data are then processed for handling missing values, removing duplicates, encoding categories, normalizing the features, selecting

features, and balancing classes prior to training the models. The processed data are then split into the training and test sets based on the mentioned 80:20 partition, with the training data being utilized for training and balancing the classes and the test set being kept aside for evaluation purposes.

3.3 Data Preprocessing and Feature Selection

The selected features of health and lifestyle are initially split into training and testing sets using the defined 80:20 setup. The training set is utilized for estimating the necessary parameters to perform data preprocessing and feature selection, whereas the testing set is kept aside for conducting an independent assessment. Such splitting ensures that there will be no leakage of the information from the testing set into the modeling procedure.

During the feature preparation in the training set, missing numerical observations are filled in using the median imputation technique, while categorical features are converted into numerical features using encoding techniques. After duplicate and inconsistent data deletion, the numerical features are standardized with the help of Z-score normalization in order to equalize their magnitudes.

After the data cleansing and transformation, the feature selection step utilizes correlation analysis and ranking of feature importance in the training dataset. After the identification of informative features, the selected representation is used to both training and testing sets of data. This process ensures that no information from the independent testing subset leaks into the feature selection process or parameter estimation.

3.4 Class Balancing and Data Partitioning

The dataset is split into training and testing datasets in the ratio of 80% and 20% respectively, making up 366,136 samples in the training dataset and 91,534 samples in the testing dataset as per the configuration of the dataset provided in Table 1.

In order to balance the class distribution in the training dataset, Synthetic Minority Over-sampling Technique (SMOTE) is employed on the training dataset only. Synthetic Minority Over-sampling Technique creates synthetic instances of minority classes based on interpolation between an instance of a minority class and one of its nearest neighbors. The synthetic sample is expressed as:

$$X_{\text{new}} = X_i + \lambda (X_{nn} - X_i) \quad (1)$$

where X_i represents an existing minority-class sample, X_{nn} denotes a neighbouring minority-class sample, and λ is a randomly generated interpolation factor in the range $[0, 1]$. The resulting balanced training set is then supplied to the machine learning models for training, whereas the original testing subset remains unchanged for independent performance evaluation.

3.5 Machine Learning Model Development

After class balancing, five supervised machine learning techniques are generated and compared for disease-risk prediction: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, and Random Forest. The algorithms are then trained using the balanced training dataset and then tested on an independent testing set. This configuration allows to examine the prediction ability of various classification techniques in a unified feature set.

Among the evaluated models, the Random Forest algorithm is selected as the primary predictive model since the process involves classification of individuals by considering various demographics, health-related factors, and lifestyles. The Random Forest algorithm builds a forest of decision trees from bootstrap samples. The quality of a candidate tree split is measured using the Gini impurity:

$$G = 1 - \sum_{c=1}^C p_c^2 \quad (2)$$

where C denotes the number of classes and p_c represents the proportion of samples belonging to class c . Lower impurity indicates a more homogeneous node. The structure of classification learned by each tree is based on a distinct sample of training data, and this combination of individual classifications yields the final classification result.

3.6 Risk Prediction, Explainability and Recommendation Generation

After a comparative evaluation, the selected Random Forest algorithm is applied to create the final disease-risk prediction using the input features. The output of the model is calculated

by voting and is denoted as follows:

$$\hat{Y} = \text{mode} \{T_1(X), T_2(X), \dots, T_K(X)\} \quad (3)$$

where $T_k(X)$ represents the prediction produced by the k -th decision tree for the input feature vector X , and $\hat{Y}$ denotes the final predicted class. The ensemble decision rule decides about the prediction from among Healthy, Diabetes Risk, Hypertension Risk, Obesity Risk, Heart Disease Risk, and Kidney Disease Risk.

The explainability module utilizes the inherent feature-importance functionality of the trained Random Forest algorithm. The feature-importance values are analyzed to obtain the relative importance of the input features for the classification task. This gives an interpretation of the relationship between the output and the demographic, health, and lifestyle factors being used for predictions. Age, frequency of smoking, and sleeping time are some of the important features found in the analysis. Therefore, this framework incorporates model comparisons with interpretable prediction using the Random Forest classifier, and not merely using the five classifiers for performance comparisons.

The predicted category is then passed to the recommendation module. Recommendations for prevention based on the risk category are offered in terms of nutrition, exercise, stress reduction, sleep, and consultation with the physician.

4. Results and Discussion

4.1 Experimental Setup and Evaluation Criteria

The experiment has been performed using Python 3.10 and Scikit-learn. The data set has been split into training and testing sets in the proportion of 80:20. The training set has been used for training of classifiers and the testing set was left aside to measure the prediction performance of trained classifiers.

Five machine learning algorithms – Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Random Forest have been studied with the same experiment configuration. The following criteria were used to measure the classifier performance: accuracy, precision, recall, and F1-score. In the case of Random Forest model, the additional

criterion of specificity has been used.

4.2 Comparative Performance of Machine Learning Models

Predictive performance of the five classifier algorithms is summarized in Table 2. The Random Forest classifier demonstrated the best classification accuracy in the test among the set of classifiers, whereas Decision Tree, SVM, KNN, and Logistic Regression were ranked second through fifth accordingly. Better performance of the ensemble algorithm reveals that using a number of decision trees is more appropriate to reflect the interdependencies between the selected health and lifestyle attributes.

Table 2. Comparative accuracy of the evaluated machine learning models

Algorithm	Accuracy (%)
Logistic Regression	84.6
SVM	87.9
KNN	85.7
Decision Tree	89.4
Random Forest	94.8

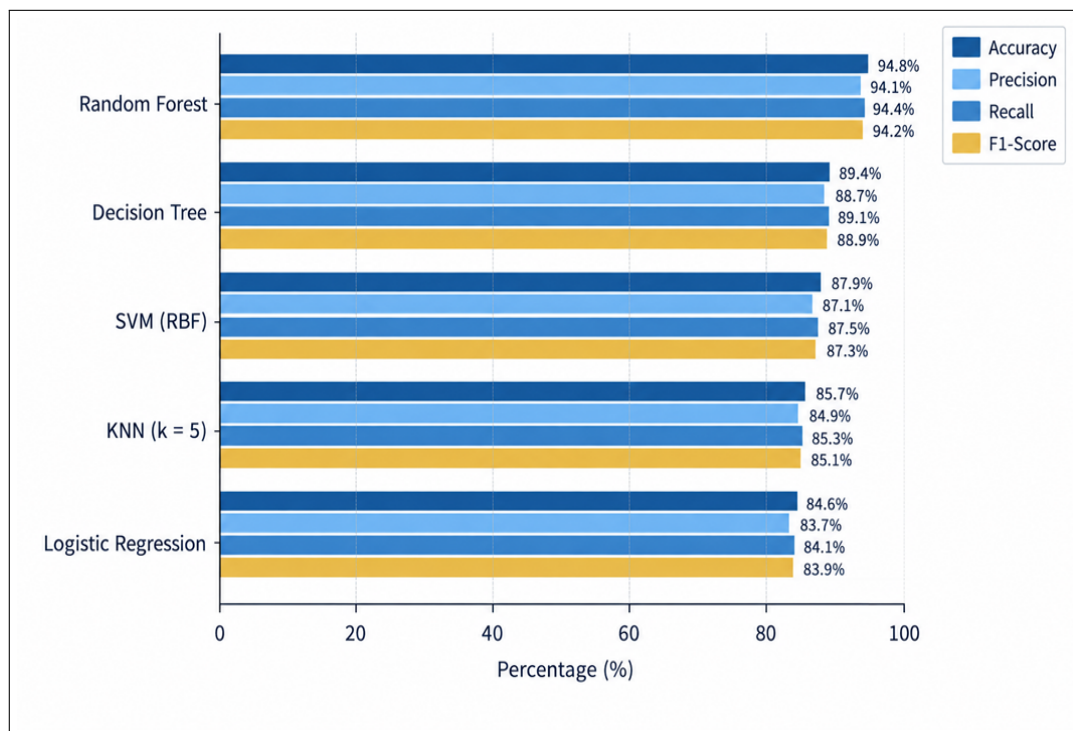


Figure 2. Comparative performance of the evaluated machine learning models

As can be seen in Figure 2, the results have shown that tree-based models outperform linear and distance models in terms of classification capabilities in the experiment performed. Random Forest gives the highest accuracy, which justifies the choice of this model as the main prediction tool for further risk analysis. Being an ensemble method, it allows using several decision trees for classification.

4.3 Feature Importance and Model Interpretability

An importance assessment was made to determine the role of the input features in the process of the Random Forest Classification. An importance assessment is done in order to have a more interpretable view of the features considered in the selected model for creating the risk classification of the disease.

These features provide an explanation on how the output of the prediction is linked with the input demographic and lifestyle features. Age is a demographic feature, while smoking habits and sleep time are lifestyle-related features. Thus, the feature importance ranking provides more information on how the classification process is carried out and helps in understanding which features contribute to the risk profile.

4.4 Risk Prediction and Recommendation Analysis

The trained Random Forest classifier is then incorporated into the prediction phase to handle the processed attributes of health and lifestyle factors in order to classify the disease risk. Prediction uses the same representation of features as the one used in developing the model to ensure uniformity in input data in training and application phases. The classification result is fed back into the application UI layer for user presentation. After the risk classification phase, the recommendations part maps the classified risk into predetermined preventive suggestions. These recommendations take into consideration lifestyle aspects such as diet, physical activities, stress handling, sleep, and doctor visits. This phase transforms the classification problem into the risk assessment problem in the application context where the predictions come along with the preventive suggestions instead of a definite clinical diagnosis.

5. Conclusion and Future Work

This research work proposed a machine learning framework to support the screening of risk factors in lifestyle diseases based on the integration of health and behavior attributes along

with the steps of preprocessing, feature selection, class balancing, comparison of classifiers and explainability of predictions. The application of Random Forest algorithm will be used for prediction purposes, and feature importance will provide an insight into the health and lifestyle factors that influence the disease risk. The framework incorporates user interaction, data processing, predictions, persistence and recommendations into one application while keeping in mind the differentiation between the automated screening and the clinical diagnosis process. Future research may include wearable and sensor-generated health information, longitudinal studies, advanced explainability techniques, adaptive recommendations, population studies and incorporation into healthcare platforms for ongoing risk assessment.

References

- [1] World Health Organization. Noncommunicable diseases progress monitor. World Health Organization, 2025.
- [2] Mensah, George A., Gregory A. Roth, and Valentin Fuster. "The global burden of cardiovascular diseases and risk factors: 2020 and beyond." *Journal of the American college of cardiology*, 2019, vol. 74, no. 20: 2529-2532.
- [3] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [4] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." *BMC Medical Informatics and Decision Making*, 2019, vol. 19, no. 1: 281.
- [5] Battineni, Gopi, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. "Applications of machine learning predictive models in the chronic disease diagnosis." *Journal of Personalized Medicine*, 2020, vol. 10, no. 2: 21.
- [6] Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A next-generation hyperparameter optimization framework." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, 2623-2631.

- [7] Hossain, Khondoker Murad, and Tim Oates. "Ten-Guard: tensor decomposition for backdoor attack detection in deep neural networks." arXiv preprint, 2024, arXiv:2401.05432.
- [8] Sumalatha, P., M. Akanksha, and Maria Fatima. "Lifestyle Disease Prediction." International Journal of Scientific Research in Science, Engineering and Technology, 2023, vol. 10, no. 3: 39-42.
- [9] Arumugam, K., Mohd Naved, Priyanka P. Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings, 2023, vol. 80: 3682-3685.
- [10] Tiwari, Achyut, Aryan Chugh, and Aman Sharma. "Ensemble framework for cardiovascular disease prediction." Computers in Biology and Medicine, 2022, vol. 146: 105624.
- [11] Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System (BRFSS): 2024 Survey Data and Documentation, Available: https://www.cdc.gov/brfss/annual_data/annual_2024.html