Health Indicator Predictions from lifestyle and biometric data using Machine Learning Models

Manuela Pop

University of Texas at Austin, MSAI

Table of Contents

Abstract

This study investigates the ability of machine learning models, specifically neural networks and tree-based classification, to predict the likelihood of being healthy versus having a disease using a dataset comprising 100,000 surveyed individuals, based on lifestyle, biometric, behavioral, and demographic factors.

Despite applying feature scaling and data preprocessing to the dataset, the models were unable to predict with high accuracy whether individuals were healthy or diseased based on the input features provided. The findings accentuate the importance of rich and comprehensible feature input and effective data integration in enhancing prediction accuracy. These results have further been compared to existing studies, such as those by Kim et al. (2024) and Effiok et al. (2022), which link predictive models to real patient data, demonstrating that real-world scenarios require richer, more diverse, and comprehensive input data.

Health Indicator Predictions from lifestyle and biometric data using Machine Learning Models

Introduction and Research Background

A well-known English proverb says that eating an apple a day can help keep the doctor away. What does this proverb mean? Is it enough to eat a balanced diet rich in fruits and vegetables to be healthy, or do we also need to incorporate other doctor-recommended practices, such as regular exercise and good sleep habits? Researchers have found that a vegetarian or Mediterranean diet is most suitable for promoting longevity. Additionally, other guidelines include ensuring we reach our 10,000 daily steps, get at least eight hours a sleep a night, drink eight cups of water per day, reduce our sunlight exposure and wear sunscreen, manage our weight and stress, and reduce our screen time. Several other recommendations include getting pets, maintaining quality friendships, and building a network of support. The list can go on and on about the best lifestyle choices we have been advised to follow for ensuring longevity and a healthy lifestyle. Each day, new studies are published on what we should eat, how much we should exercise, and other factors that can influence our longevity and health. Shall we believe everything we read, or shall we take everything with a grain of salt and a glass of wine?

Numerous studies have gathered biometric data and analyzed it to conduct research and predict longevity. Innumerable books have been written, and researchers and the medical community have traveled across the globe to discover the secret spots where individuals live the longest and are the healthiest, researching those communities to uncover their secrets to longevity. Such books have been written about the 5 “Blue Zones”, where people live the longest and have the best health. These “Blue Zones” are spread in different parts of the world, unrelated to one another. They are in Okinawa (Japan), Sardinia (Italy), Ikaria (Greece), Loma Linda, California (North America), and the Nicoya Peninsula in Costa Rica (Central America). The only relations they have are the results gathered. Several factors have been examined in these studies, including access to community resources, exercise, diet, and alcohol consumption. The resulting data showed that a Mediterranean diet could be one of the best to follow to ensure longevity, together with fostering a sense of community and friendship, promoting exercise, hobbies, and a positive attitude. These have also always been general guidelines given by doctors to the population.

Most of us tend to be conscious about our lifestyle options, exercise, and diet, and follow some of these guidelines. We shall investigate whether we could substantiate any of these theories related to lifestyle choices and their impact on health.

Motivated by current studies using machine learning, extensive published literature, and a comprehensive database of doctor notes and guidelines, the purpose of this study was to determine whether it is possible to infer an individual's health status using machine learning models, such as neural networks and tree-based classification, based on a provided survey-based dataset that encompasses multiple lifestyle and biometrics features, gathered from 100,000 individuals.

Research and Methods

This study investigates the ability of machine learning models, specifically neural networks and tree-based classification, to predict the likelihood of being healthy versus having a disease using a dataset comprising 100,000 surveyed individuals, based on lifestyle, biometric, behavioral, and demographic factors. Some of the input features examined can or cannot prove whether the doctor’s guidelines are proper.

The data set analyzed in this study was downloaded from Kaggle and includes features such as BMI, blood pressure, heart rate, cholesterol, glucose, insulin, physical activity, hours of sleep, smoking status, sugar levels, exercise, sunlight exposure, and overall health status, indicating whether the individual is healthy or has a disease. A combination of risk factors has been identified as affecting the likelihood of disease progression, including stress levels and sleep duration. Survey-based data was analyzed in this experiment by training models and plotting the results.

This experiment aims to evaluate survey-based data using machine learning models, such as recurrent neural networks (RNNs) and a tree-based classification model optimized for large datasets, to determine whether we can predict disease risk based on lifestyle feature indicators included in the dataset. In recent years, machine learning models have been utilized to analyze large clinical datasets, employing nonlinear and complex analysis to predict patterns and understand how this data can impact our health or predict disease. These sophisticated techniques can provide researchers, the medical community, and the general population with deep insights and support after the results are interpreted.

Can an ML model accurately learn to identify the boundary between healthy and diseased individuals, given a set of features such as BMI, cholesterol, heart rate, and diet, based on surveyed data? Can an ML model offer a more insightful interpretation of this data to researchers and the medical community? Do model types, tree-based versus neural, perform differently, given the limitations of the input features? How do the evaluation metrics and plots interpret the results and reveal the limitations and errors in the predictions? How do these results compare to those of current studies, considering that the models were trained using different datasets gathered from a completely different population set? Can these results relate at all, or do they fail to interpret the data in a manner that indicates whether an adult can improve their lifestyle choices to achieve a healthier well-being? Do these results reflect any of the vast studies performed by well-known researchers who traveled across the globe in search of the "secret elixir of life"?

Despite applying feature scaling and data preprocessing to the dataset, the models were unable to predict with high accuracy whether individuals were healthy or diseased based on the input features provided. The findings accentuate the importance of rich and comprehensible feature input and effective data integration in enhancing prediction accuracy. These results have further been compared to existing studies, such as those by Kim et al. (2024) and Effiok et al. (2022), which link predictive models to real patient data, demonstrating that real-world scenarios require richer, more diverse, and comprehensive input data.

Materials and Data Sources

The dataset used is health_lifestyle_classification.csv, which was downloaded from the Kaggle datasets website, and contains 40 biometric and lifestyle feature inputs, including demographic, biological, and behavioral data from 100,000 surveyed individuals. These columns are split into two categories: numerical features and categorical features. The numerical feature columns are age, BMI, blood pressure, cholesterol, heart rate, glucose, insulin, calorie intake, sugar intake, screen time, stress level, mental health score, and exercise training hours. The categorical feature columns are gender, marital status, diet type, occupation, sleep quality,

mental health support, exercise type, device usage, healthcare access, insurance, family history, sunlight exposure, pet owner, caffeine intake, and meals per day. The last column is the target, binary classification: “healthy vs diseased”. The target column indicates whether an individual has been diagnosed with a disease or not. The classification is based on a combination of medical and lifestyle indicators collected from the individual.

https://www.kaggle.com/datasets/mahdimashayekhi/disease-risk-from-daily-habits/data

Author: Mahdi Mashayekhi Usability Score: 10.0

As mentioned on the website, this data has been collected and aggregated from multiple anonymized sources, such as:

- Wearable health devices
- Survey forms
- Fitness tracking apps
- Self-reported medical questionnaires
- Health monitoring programs under voluntary participation

Figure 1 shows the data distribution, with 70,000 records for Healthy and 30,000 records for Diseased. Data preprocessing involved handling missing values, encoding categorical features using one-hot encoding, and normalizing data using MinMax scaling. The dataset was partitioned into 80/20 training and test split subsets. Exploratory data analysis revealed class imbalance (~70% healthy, ~30% diseased). Feature correlations with the target were near zero, indicating weak linear associations. These observations underscore the need for nonlinear models to capture the complex interactions among the features.

The methods used were two learning models:

1. Recurrent Neural Network (RNN) – Keras SimpleRNN with Adam Optimizer (learning rate r=0.001), 20 epochs, batch size 256. The Model has an accuracy of 0.7, with an F1-score of 0.824.
2. A gradient-boosted decision tree-based classification model (learning rate=0.08, max iterations of 500, early stopping= true, random state=seed). The model has an accuracy of 0.70, with an F1-score of 0.824.

Results

The confusion matrix, Figure 2 RNN, shows True healthy = 14014 (TN), True Diseased = 5980 (FN), 5 FP, and 1TP. Nearly all Diseased individuals were incorrectly predicted as healthy, with only five people being wrongly predicted as healthy, and one healthy person was incorrectly predicted as diseased.

The confusion matrix for the tree-based classifier (Figure 3) shows that True healthy = 14,019, predicted healthy, and True Diseased = 5,981, wrongly predicted healthy, when the threshold is set to 50. With a threshold of 15 (Figure 4), the tree-based classifier shows True Healthy = 14,019, false positives, and True Diseased = 5,981, who were indeed expected to be diseased.

The models cannot distinguish between healthy and diseased individuals with high accuracy, suggesting that predictions based on the survey data remain speculative without richer and stronger features from a wider clinical research dataset. Based on the Confusion Matrix formula calculations, the overall accuracy is 0.7, which corroborates the model accuracy results. The recall is 0.017%, the precision is 17%, and the specificity is 99%, which indicates that the model correctly identifies healthy individuals but fails to identify diseased individuals, resulting in class imbalance. This is primarily the case when working with a dataset where one class significantly outweighs the other, as in this case, where Healthy has 14,014 and Diseased has 5,980.

Figures 11 and 12 display the precision-recall curves for the models used. The lines are mostly flat along the x-axis, indicating low precision and poor discrimination, which leads to false positives, as also shown by the confusion matrix.

Figures 5 and 6 show the RNN-predicted probability of being healthy versus diseased for the training and test data results. X-axis predicted probability of an individual to be diseased, Y-axis the count of individuals with the probability of being diseased: blue bars, healthy individuals; orange bars, diseased individuals. There is no overfitting, but also no meaningful learning. The model assigns the same predicted probability, 0.3, at the peaks. As with the confusion matrix, these graphs indicate that the models cannot accurately predict whether an individual is healthy or diseased based on the cumulative input features provided in the dataset.

Figures 7 and 8 show the false positive rate on the x-axis and the actual positive rate on the y-axis. The AUC is 0.5, which represents random guessing. Since the ROC shows a diagonal line with no curve, the results indicate that both models do not perform optimally, have insufficient feature inputs, and are unable to effectively discriminate between healthy and diseased individuals. When the feature inputs are weak, the models cannot extract meaningful patterns.

Feature Importance interpretation:

Figure 9 illustrates the feature importance, with work hours and daily steps being the most significant, followed by water intake, cholesterol, income, BMI, sleep hours, and screen time, among others.

Figure 10 illustrates the Shapley-based feature value, explaining how SHAP values correlate to feature values and the impact of the model output. The y-axis displays the feature importance from top to bottom. The x-axis represents the SHAP value, indicating the degree of change. The color of each point on the graph represents the value of the corresponding feature, with red indicating high values and blue low values. Each point represents a row of data in the dataset. Positive SHAP values indicate an increase in the predicted probability of disease, while negative values indicate healthy predictions. The red dots to the left of 0 are associated with a healthier prediction (i.e., a lower probability of the disease). The results in this experiment demonstrate how the features appear in order of their sum of SHAP values. It illustrates how the feature influences the model's output, resulting in either a higher or lower prediction. For example, work hours were the most influential feature.

High cholesterol increases the likelihood of the disease. The same applies to BMI, stress, weight, glucose levels, insulin levels, and blood pressure; the higher the value, the higher the

association with an increased probability of disease. Increased exposure to sunlight is also associated with a higher likelihood of the disease.

Meals per day, exercise type, and calorie intake had no significant effect on model prediction. Contrary to what is commonly known about diets, the vegetarian diet did not influence the prediction in this data set. The sampled population may not have enough vegetarian data to influence this study. Similarly, the omnivore diet had little effect on the model's prediction of disease status versus health. According to this study, eating an apple a day may not have a significant influence on keeping the doctor away. Neither will a Mediterranean diet. However, this dataset lacks sufficient samples from the population to establish whether a diet can have a significant influence on health. If we were to gather data from the 5 “Blue Zones” together with the same amount of data from the rest of the world, would we see different results, and could the weight of the diet feature values influence a healthier prediction? After all, enough scientific evidence has been gathered to prove that a diet full of fruits and vegetables is much healthier than one full of fast food. Perhaps this data set used in this experiment is not rich enough to demonstrate whether diet can have any effect on model prediction in healthy versus diseased individuals.

Interestingly, the occupation of a teacher in this dataset is associated with a higher likelihood of being in good health. The same applies to job types in the office, as well as being a doctor.

Low device usage, thereby reducing screen time, indeed increases the likelihood of being healthy.

Smoking affects the likelihood of being diseased, which is linked with what we already know so far about smoking.

Pet owners, on the other hand, don’t have a whole lot of influence on the model production.

Higher daily step counts are associated with a healthier prediction. The same applies to water intake and daily supplements. Hence, achieving 10,000 steps is worth it and can significantly improve our health. Similarly, 8 cups of water do indeed contribute to a healthier status.

However, in this dataset, the physical activity reveals a higher prediction of the disease, which doesn’t make sense. Perhaps the data is not accurate, or other factors are at play that

outweigh the prediction, such as the possibility that someone who is physically active may have higher cholesterol, a higher intake of sugar, or higher glucose levels. Maybe some individuals already have existing medical conditions and are trying to adopt better lifestyle choices to improve their health. Alternatively, this can also indicate that people who work in physically demanding jobs, such as construction or mining, or maybe they are professional athletes who can exert themselves and potentially injure themselves, thereby increasing the likelihood of being classified as diseased. This data set is not sufficiently indicative or comprehensive for us to draw solid conclusions based on medical evidence, as we lack knowledge of the context behind all this data. We don't know where this data set originates, which part of the world it represents, or to which population it pertains.

In any case, as current studies show, we need a richer, more diverse, and a much larger dataset with a lot more complex features to accurately predict healthy versus diseased individuals based on lifestyle and biometric factors. The SHAP feature values graph represents the best indicator of how the model predicted diseased vs healthy based on these features.

Discussion and Conclusion

This experiment demonstrates that while lifestyle feature inputs provide a moderate foundation for health classification, they are collectively insufficient for a robust prediction of healthy versus diseased individuals when used in isolation as model training/test parameters.

It was interesting to analyze the importance of each feature in model prediction and see how it affects the results. The SHAP value graph provides most of the insights into this study. The values support some of the theories we have been told about our lifestyles, as analyzed in the results section above. These resulting feature values corroborate with multiple existing findings on lifestyle and biometric factors identified by researchers as influencing a healthier lifestyle. However, considering the entire dataset collectively, the model failed to accurately predict whether an individual is healthy or diseased.

This study corroborates Kim et al.'s (2024) research, which suggests that a model's success depends on how well the input data captures underlying physiological processes. Kim et al.'s (2024) study, based on data gathered from middle-aged South Korean adults, utilized machine learning to predict the quality of life, rather than distinguishing between healthy and diseased individuals, and found that stress and sleep quality were the top predictive factors. The

articles reinforce the conclusion from this study that behavioral/lifestyle features are valuable but often insufficient by themselves for distinguishing between Healthy and Diseased individuals in cross-sectional data when training machine learning models and predicting outputs. Similarly, as in Kim et al.'s (2024) study (https://pmc.ncbi.nlm.nih.gov/articles/PMC10785386), the SHAP values indicate that stress, BMI, and activity have a significant impact on predicting disease. Like Kim et al.'s model performance, the F1-score of 0.72 from the dataset and classification model used in their study is comparable to the range found in this study, which is 0.82. However, one can say that the F1-score using the RNN in this study is higher and hence better than Kim et al.'s study. One difference is that Kim et al.'s study found sleep quality to be a major predictor in healthy individuals.

In contrast, this study suggests that the number of sleep hours has no significant influence on the model prediction. Although there is a clear distinction between the number of sleep hours and sleep quality, one can still achieve excellent sleep quality even if they sleep fewer hours. Kim et al.'s study claims success with a F1-score of 0.72. As we have learned from machine model training, an F1-score close to 1 is considered much more successful. However, Kim et al. (2024) achieved a meaningful separation, whereas the results in this experiment show poor separation. Nevertheless, the results in this experiment provide valuable insights into feature analysis and how each of these lifestyle factors and biometrics can influence the model's prediction of healthy vs. diseased individuals.

Effiok et al. (2022) (https://pmc.ncbi.nlm.nih.gov/articles/PMC9160810/) focused their work on modeling the cumulative effect of many lifestyle risk factors on complex diseases. They collected risk factors into preventive, permissive, and core categories and proved that combined exposure to unhealthy behaviors increases disease risk. They studied the accumulation of risk factors and had stronger performance, rather than predicting healthy versus diseased individuals. This study's approach included multiple input feature factors; however, it lacked the clinical depth that the Effiok et al. study encompassed. The results of this analysis showed that using a simpler survey data and calculation models, one cannot achieve the same level of complexity and accuracy in the results.

While prior studies achieved meaningful prediction of disease using integrated lifestyle and biomarker datasets, those analyses benefited from richer clinically validated features. The dataset used in this study, limited to survey-style feature indicators, lacked the same

physiological depth. The articles used to compare the results in this study used larger and more comprehensive datasets, including biomarkers and interaction modeling. The signals from this data set are weak if the features do not come from a richer clinical background. In conclusion, the two peer studies had a higher and more accurate prediction and better results explanations based on richer and comprehensible clinical feature inputs. In the end, although less successful than theirs, the analysis of this study demonstrates that it is not easy to predict whether a sample is healthy or diseased based on a simpler survey-based data. Yet, it provides insight by evaluating the lifestyle input feature values and explaining how they influence model predictions.

References

1. Kempen GI, Ormel J, Brilman EI, Relyveld J. Adaptive responses among Dutch elderly: the impact of eight chronic medical conditions on health-related quality of life. Am J Public Health. 1997;87(1):38-44. doi:10.2105/AJPH.87.1.38.
2. Martinez R, Lloyd-Sherlock P, Soliz P, et al. Trends in premature avertable mortality from non-communicable diseases for 195 countries and territories, 1990–2017: a population-based study. Lancet Glob Health. 2020;8(4):e511-e523. doi:10.1016/S2214-109X(20)30035-8.
3. World Health Organization. Global Health Estimates: life expectancy and leading causes of death and disability. Geneva: WHO; 2020. https://www.who.int/data/gho/data/themes/theme-details/GHO/mortality-and-global-health-estimates. Accessed Nov 11, 2025.
4. Bowling A, Dieppe P. What is successful ageing and who should define it? BMJ. 2005;331(7531):1548-1551. doi:10.1136/bmj.331.7531.1548.
5. Giannouli P, Zervas I, Armeni E, et al. Determinants of quality of life in Greek middle-aged women: a population survey. Maturitas. 2012;71(2):154-161. doi:10.1016/j.maturitas.2011.11.013.
6. Makovski TT, Schmitz S, Zeegers MP, Stranges S, van den Akker M. Multimorbidity and quality of life: systematic literature review and meta-analysis. Ageing Res Rev. 2019;53:100903. doi:10.1016/j.arr.2019.04.005.
7. Tan SL, Storm V, Reinwand DA, Wienert J, de Vries H, Lippke S. Understanding the positive associations of sleep, physical activity, fruit and vegetable intake as predictors of quality of life and subjective health across age groups: a theory-based, cross-sectional web-based study. Front Psychol. 2018;9:977. doi:10.3389/fpsyg.2018.00977.
8. Lidin M, Ekblom-Bak E, Rydell Karlsson M, Hellénius ML. Long-term effects of a Swedish lifestyle intervention programme on lifestyle habits and quality of life in people with increased cardiovascular risk. Scand J Public Health. 2018;46(6):613-622. doi:10.1177/1403494817746536.
9. Marcos-Delgado A, Hernández-Segura N, Fernández-Villa T, Molina AJ, Martín V. The effect of lifestyle intervention on health-related quality of life in adults with metabolic

syndrome: a meta-analysis. Int J Environ Res Public Health. 2021;18(3):887. doi:10.3390/ijerph18030887.

10. Colpani V, Baena CP, Jaspers L, et al. Lifestyle factors, cardiovascular disease and all-cause mortality in middle-aged and elderly women: a systematic review and meta-analysis. Eur J Epidemiol. 2018;33(9):831-845. doi:10.1007/s10654-018-0374-z.
11. Phyo AZZ, Freak-Poli R, Craig H, et al. Quality of life and mortality in the general population: a systematic review and meta-analysis. BMC Public Health. 2020;20(1):159. doi:10.1186/s12889-020-09639-9.
12. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944.
13. Lee SK, Youn-Jung S, Kim J, et al. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. Healthc Inform Res. 2014;20(2):125-134. doi:10.4258/hir.2014.20.2.125.
14. Hatton CM, Paton LW, McMillan D, et al. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalized mental healthcare. J Affect Disord. 2019;246:857-860. doi:10.1016/j.jad.2018.12.095.
15. Lee SH, Choi I, Ahn WY, et al. Estimating quality of life with biomarkers among older Korean adults: a machine-learning approach. Arch Gerontol Geriatr. 2020;87:103966. doi:10.1016/j.archger.2019.103966.
16. Baek Y, Seo B, Jeong K, Yoo H, Lee S. Lifestyle, genomic types and non-communicable diseases in Korea: protocol for the Korean Medicine Daejeon Citizen Cohort study (KDCC). BMJ Open. 2020;10(4):e034499. doi:10.1136/bmjopen-2019-034499.
17. Daviglus ML, Liu K, Pirzada A, et al. Favorable cardiovascular risk profile in middle age and health-related quality of life in older age. Arch Intern Med. 2003;163(20):2460-2468. doi:10.1001/archinte.163.20.2460.
18. Grundy SM, Cleeman JI, Daniels SR, et al. Diagnosis and management of the metabolic syndrome: an AHA/NHLBI scientific statement. Circulation. 2005;112(17):2735-2752. doi:10.1161/CIRCULATIONAHA.105.169404.
19. World Health Organization. The Asia-Pacific Perspective: Redefining Obesity and Its Treatment. 2000. https://apps.who.int/iris/handle/10665/206936. Accessed Nov 11, 2025.

20. World Health Organization. WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region. 2007. https://apps.who.int/iris/handle/10665/206952. Accessed Nov 11, 2025.
21. Bae KH, Jang ES, Park K, Lee Y. Development of the questionnaire of cold-heat pattern identification based on usual symptoms: reliability and validation study. J Physiol Pathol Korean Med. 2018;32(5):341-346. doi:10.15188/kjopp.2018.10.32.5.341.
22. Ezzati M, Lopez AD, Rodgers A, Murray CJL, eds. Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. Geneva: WHO; 2004.
23. Yook SM, Park S, Moon HK, Kim K, Shim JE. Development of Korean Healthy Eating Index for adults using the Korea National Health and Nutrition Examination Survey data. J Nutr Health. 2015;48(5):419-428. doi:10.4163/jnh.2015.48.5.41922.
24. Sohn SI, Kim HD, Lee MY, Cho YW. Reliability and validity of the Korean version of the Pittsburgh Sleep Quality Index. Sleep Breath. 2012;16(3):803-812. doi:10.1007/s11325-011-0579-9.
25. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. 1989;28(2):193-213. doi:10.1016/0165-1781(89)90047-4.
26. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). J Public Health (Berl). 2006;14(2):66-70. doi:10.1007/s10389-006-0024-x.
27. Chang S. Standardization of Collection and Measurement for Health Data. Seoul: Kyechukmunhwasa; 2000:121-159.
28. Gandek B, Ware JE, Aaronson NK, et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. J Clin Epidemiol. 1998;51(11):1171-1178. doi:10.1016/S0895-4356(98)00109-7.
29. Pezzilli R, Bini R, Fantini L, et al. Quality of life in chronic pancreatitis. World J Gastroenterol. 2006;12(39):6249-6256. doi:10.3748/wjg.v12.i39.6249.
30. Ware JE Jr. SF-36 Health Survey update. Spine. 2000;25(24):3130-3139. doi:10.1097/00007632-200012150-00008.

31. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. doi:10.1613/jair.953.
32. Wang K, Tian J, Zheng C, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. Comput Biol Med. 2021;137:104813. doi:10.1016/j.compbiomed.2021.104813.
33. Lee D, Bin S. Structure relationships for diseased and health-related quality of life in the elderly. J Korea Contents Assoc. 2011;11(1):216-224. doi:10.5392/JKCA.2011.11.1.216.
34. Miguel AQC, Tempski P, Kobayasi R, Mayer FB, Martines MA. Predictive factors of quality of life among medical students: results from a multicentric study. BMC Psychol. 2021;9(1):119. doi:10.1186/s40359-021-00534-5.
35. Ishaq A, Sadiq S, Umer M, et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access. 2021;9:39707-39716. doi:10.1109/ACCESS.2021.3064084.
36. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. The impact of oversampling with SMOTE on the performance of three classifiers in prediction of type 2 diabetes. Med Decis Making. 2016;36(1):137-144. doi:10.1177/0272989X14560647.
37. Kim J, Mun S, Lee S, Jeong K, Baek Y. Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea. BMC Public Health. 2022;22(1):1066. doi:10.1186/s12889-022-13131-x.
38. Bae KH, Lee Y, Go HY, Kim SJ, Lee SW. The relationship between cold hypersensitivity in the hands and feet and health-related quality of life in Koreans: a nationwide population survey. Evid Based Complement Alternat Med. 2019;2019:6217036. doi:10.1155/2019/6217036.
39. Chun SW, Kim W, Choi KH. Comparison between grip strength and grip strength divided by body weight in their relationship with metabolic syndrome and quality of life in the elderly. PLoS One. 2019;14(9):e0222040. doi:10.1371/journal.pone.0222040.
40. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine-learning-based clinical decision support systems: a systematic review. Appl Sci. 2021;11(11):5088. doi:10.3390/app11115088.

41. Antoniadi AM, Galvin M, Heverin M, Hardiman O, Mooney C. Prediction of caregiver quality of life in amyotrophic lateral sclerosis using explainable machine learning. Sci Rep. 2021;11(1):22166. doi:10.1038/s41598-021-91632-2.
42. Norrish A, Jackson R, Sharpe S, Skeaff C. Prostate cancer and dietary carotenoids. Am J Epidemiol. 2000;151(2):119-123. doi:10.1093/oxfordjournals.aje.a010176.
43. Food Quality Protection Act of 1996. Public Law 104-170; 1996. https://www.congress.gov/104/plaws/publ170/PLAW-104publ170.pdf. Accessed Nov 11, 2025.
44. Delgadillo J, Moreea O, Lutz W. Different people respond differently to therapy: a demonstration using patient profiling and risk stratification. Behav Res Ther. 2016;79:15-22. doi:10.1016/j.brat.2016.02.003.
45. O'Flynn N, Staniszewska S. Improving the experience of care for people using NHS services: summary of NICE guidance. BMJ. 2012;344:d6422. doi:10.1136/bmj.d6422.
46. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet. 2014;383(9921):999-1008. doi:10.1016/S0140-6736(13)61752-3.
47. Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. Boston: Pearson Addison Wesley; 2005.
48. Cooney MT, Dudina A, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol. 2009;54(14):1209-1227. doi:10.1016/j.jacc.2009.07.020.
49. Aktas MK, Ozduran V, Pothier CE, Lang R, Lauer MS. Global risk scores and exercise testing for predicting all-cause mortality in a preventive medicine program. JAMA. 2004;292(12):1462-1468. doi:10.1001/jama.292.12.1462.
50. Framingham Heart Study. History—epidemiological background. 2019. https://www.framinghamheartstudy.org/fhs-about/history/epidemiological-background/. Accessed Nov 11, 2025.
51. Dawber TR. The Framingham Study. Ann Intern Med. 1981;94(2):286. doi:10.7326/0003-4819-94-2-286_1.
52. Feinleib M. The Framingham Study: sample selection, follow-up and methods of analyses. Natl Cancer Inst Monogr. 1983;(67):20-35.

53. PHG Foundation. The Personalized Medicine Technology Landscape. Cambridge, UK; 2018.
54. IBM Watson Health. IBM Micromedex Care Delivery Evidence-Based Clinical Decision Support for Healthcare Decision-Makers. New York; 2018.
55. QResearch. QResearch Survey. https://www.qresearch.org/. Accessed Nov 11, 2025.
56. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open. 2015;5(3):e007825. doi:10.1136/bmjopen-2015-007825.
57. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013;4(2):124. doi:10.4172/2157-7420.1000124.
58. Effiok E, Liu E, Hitchcock J. Lifestyle related risk association mining. In: International Conference on Internet of Things, Embedded Systems and Communications (IINTEC). Hamammet, Tunisia; 2018.
59. Sexton K. Cumulative health risk assessment: finding new ideas and escaping from the old ones. Hum Ecol Risk Assess. 2014;21(4):934-951. doi:10.1080/10807039.2014.946346.
60. Effiok E, Liu E, Hitchcock J. Lifestyle risk association aggregation. In: Proceedings of the Fifth IEEE International Workshop on Internet of Things: Networking Applications and Technologies. Limerick, Ireland; 2019.
61. Dietrich F, List C. Probabilistic opinion pooling generalised. Part two: the premise-based approach. Soc Choice Welf. 2017;48(4):787-814.
62. SWOP—The Prostate Cancer Research Foundation. Risk calculator. 2019. http://www.prostatecancer-riskcalculator.com/. Accessed Nov 11, 2025.
63. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.
64. World Health Organization. Global Health Risks. 2009. https://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf. Accessed Nov 11, 2025.

65. Effiok E, Liu E, Hitchcock J. Predicting cumulative effect of lifestyle risk factors for disease development. Healthc Technol Lett. 2022;—. doi:10.1049/htl2.12021. PMCID: PMC9160810. (PubMed)
66. Kim J, Jeong K, Lee S, Baek Y. Machine-learning model predicting quality of life using multifaceted lifestyles in middle-aged South Korean adults: a cross-sectional study. BMC Public Health. 2024;24(1):159. doi:10.1186/s12889-023-17457-y. PMCID: PMC10785386. (PubMed)

Figures

Figure 1. Dataset distribution

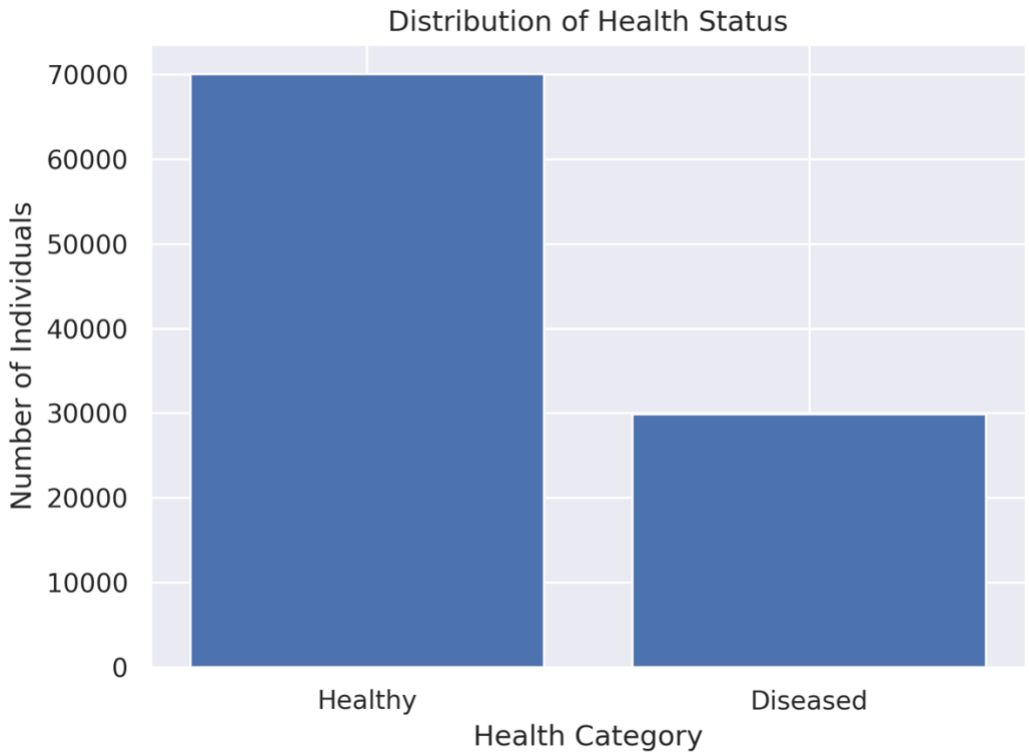

Figure 2: Confusion Matrix RNN

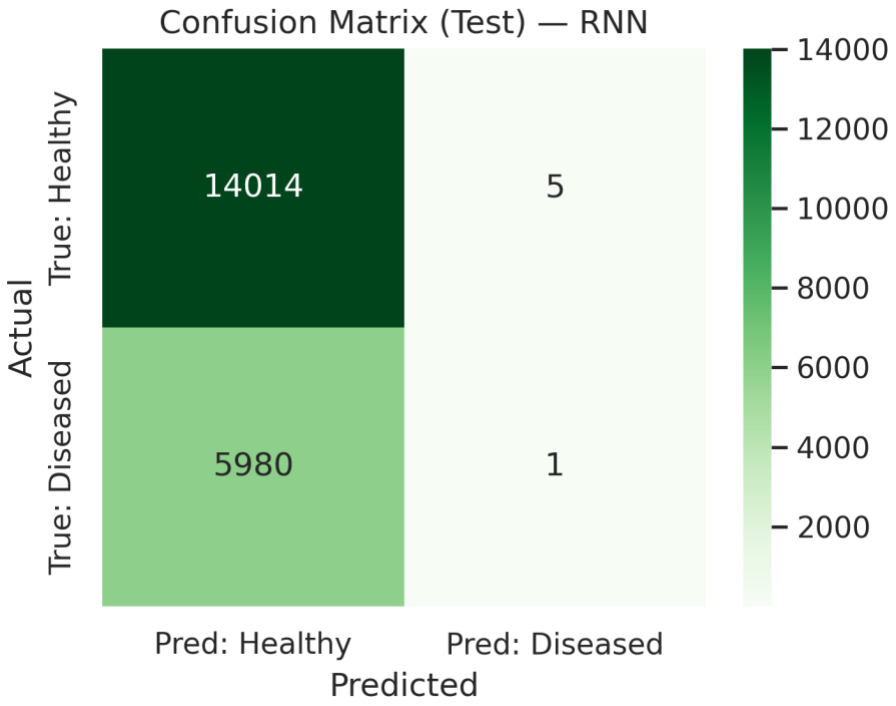

Figure 3: Confusion Matrix HGB threshold 50

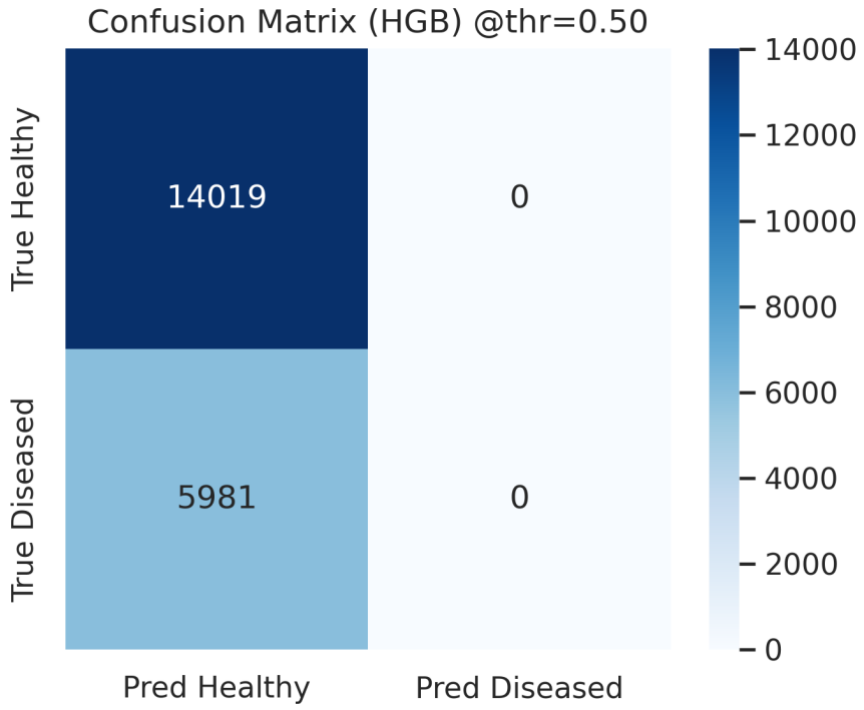

Figure 4: Confusion Matrix HGB threshold 15

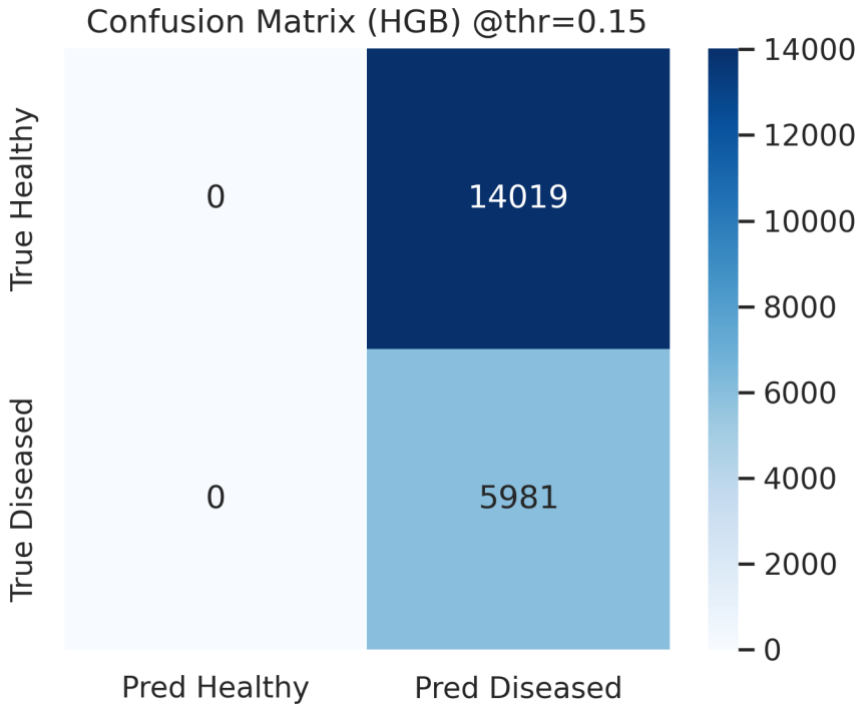

Figure 5: Test Probability Distribution - RNN

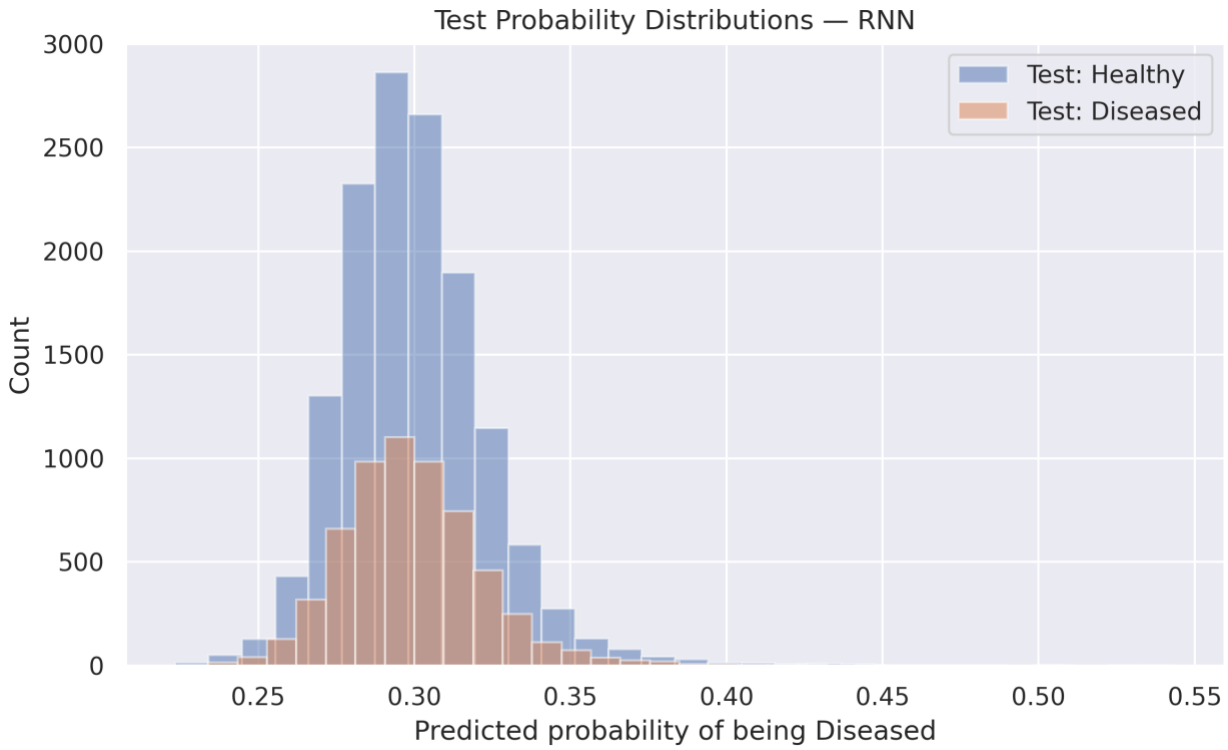

Figure 6: Train Probability Distribution - RNN

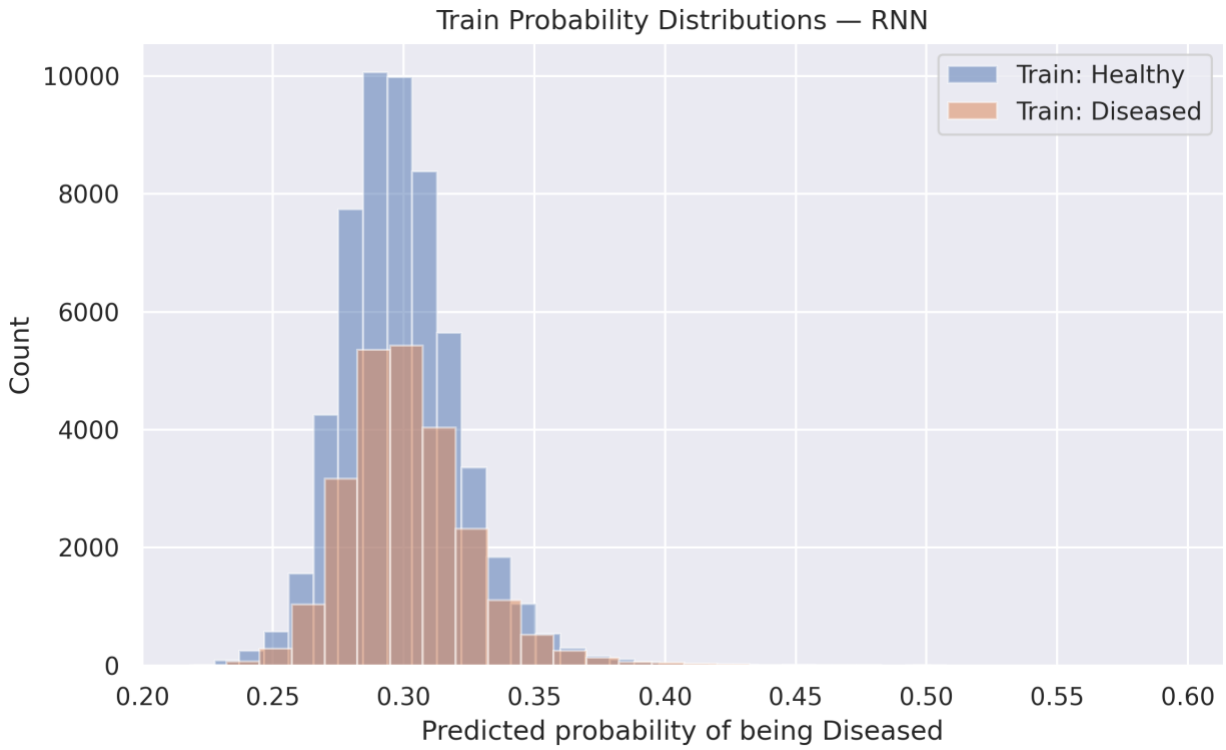

Figure 7: ROC Curve (Test Set) - R

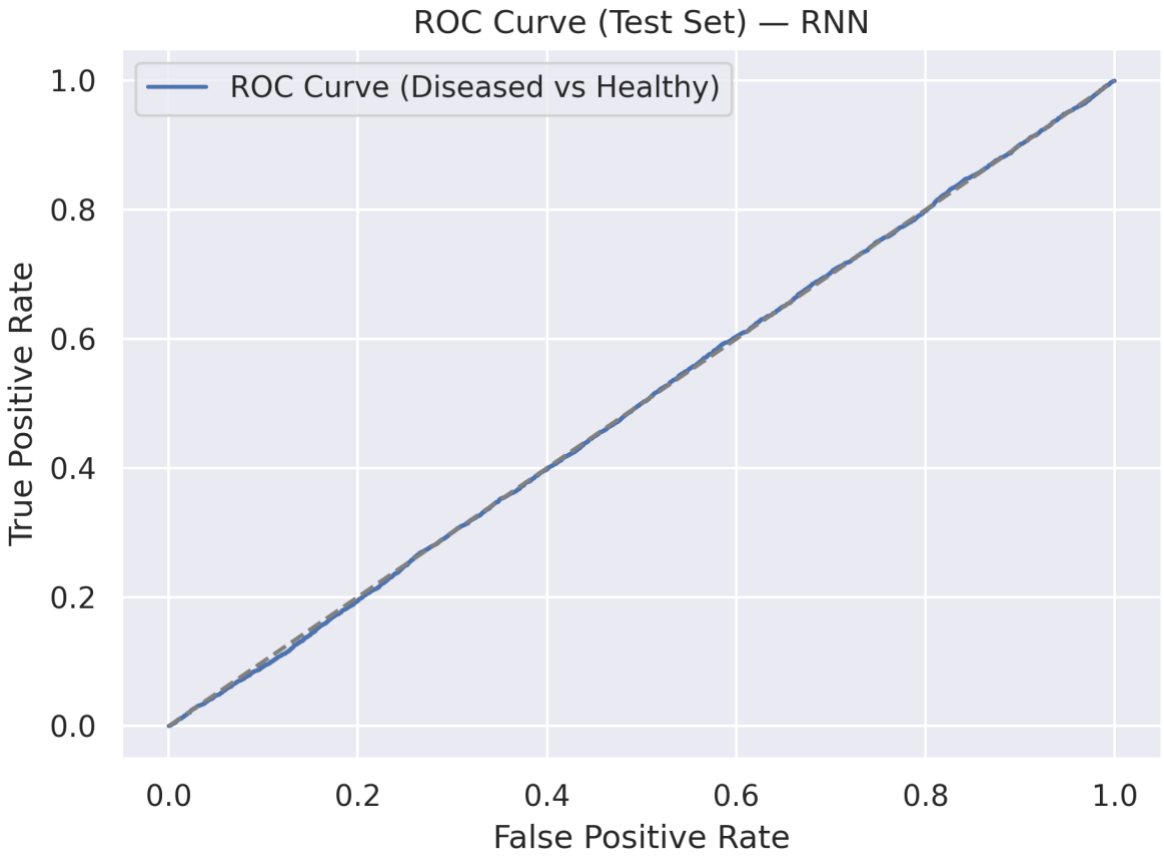

Figure 8: ROC Curve - HGB

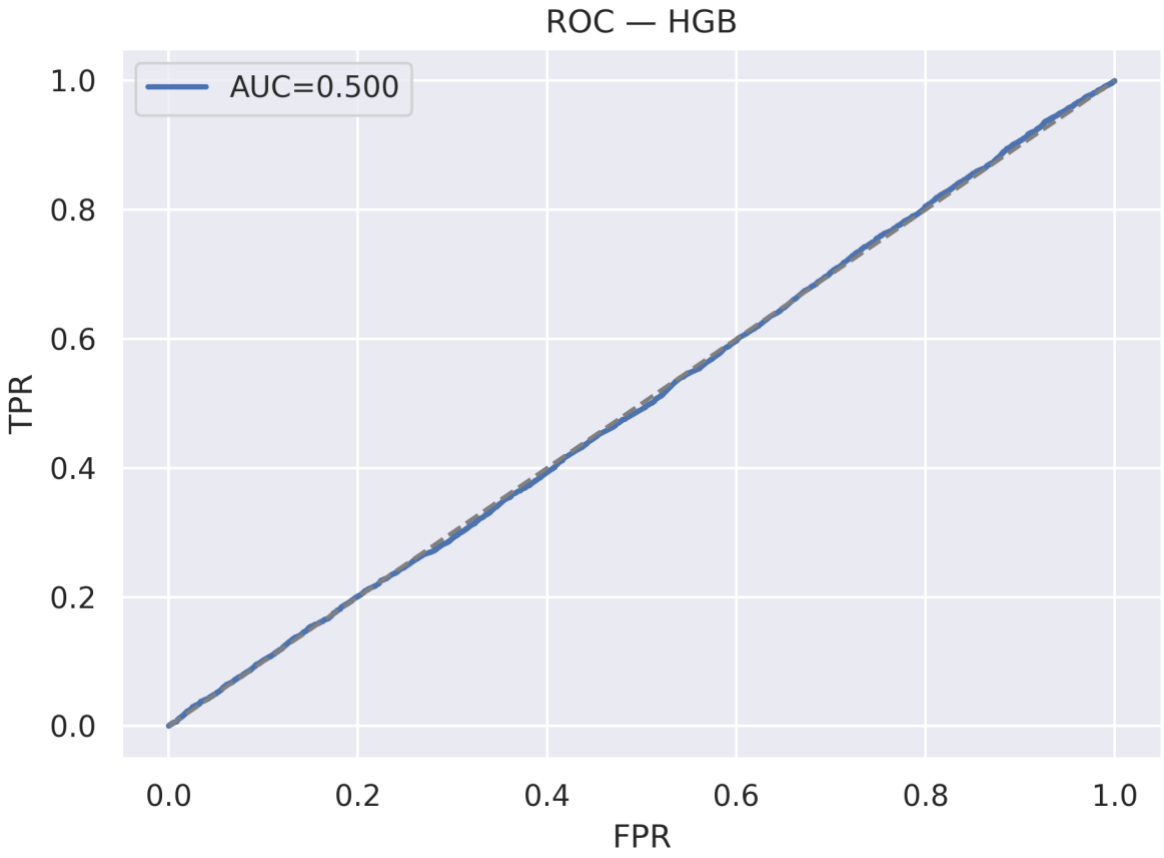

Figure 9: Feature Importance

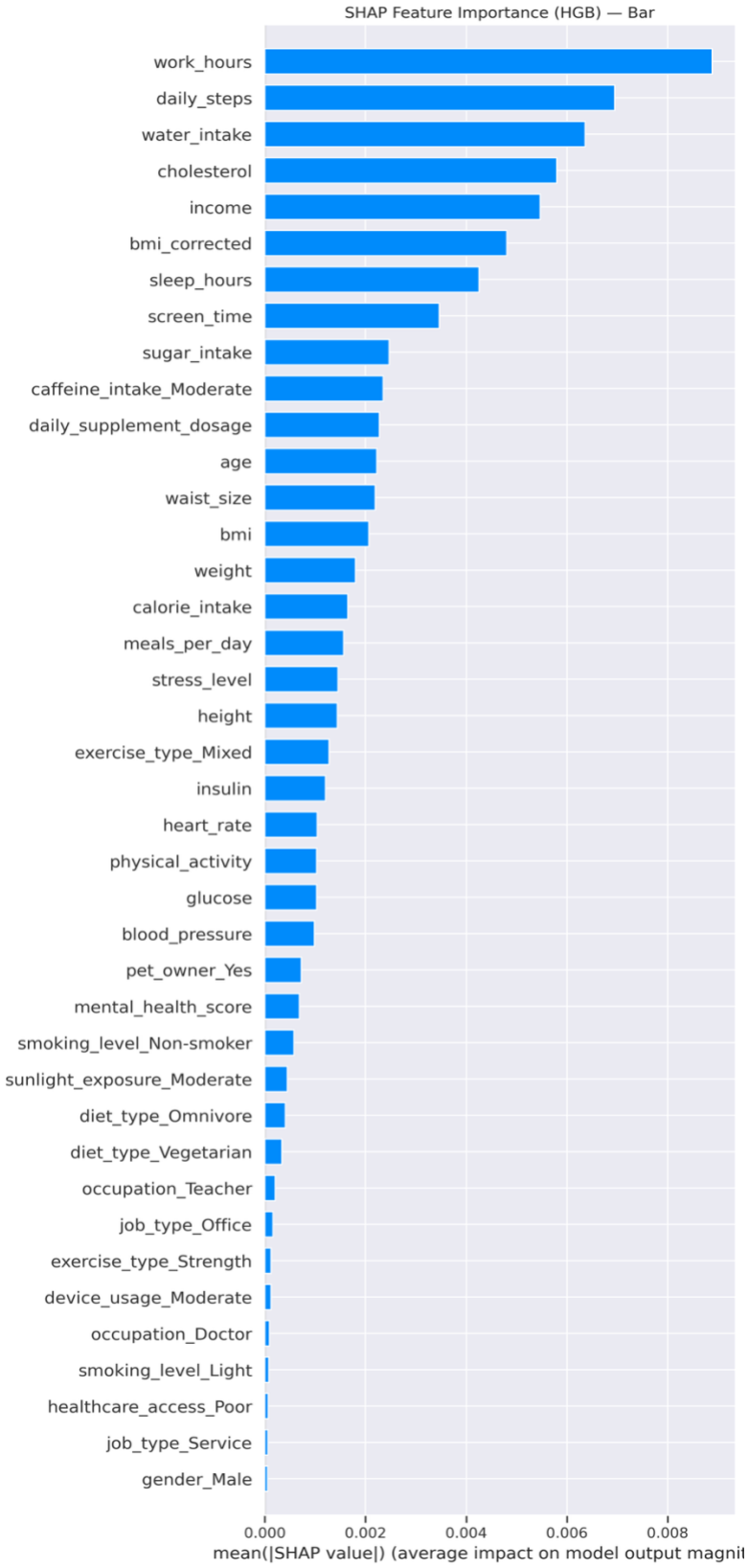

Figure 10: SHAP Feature Importance

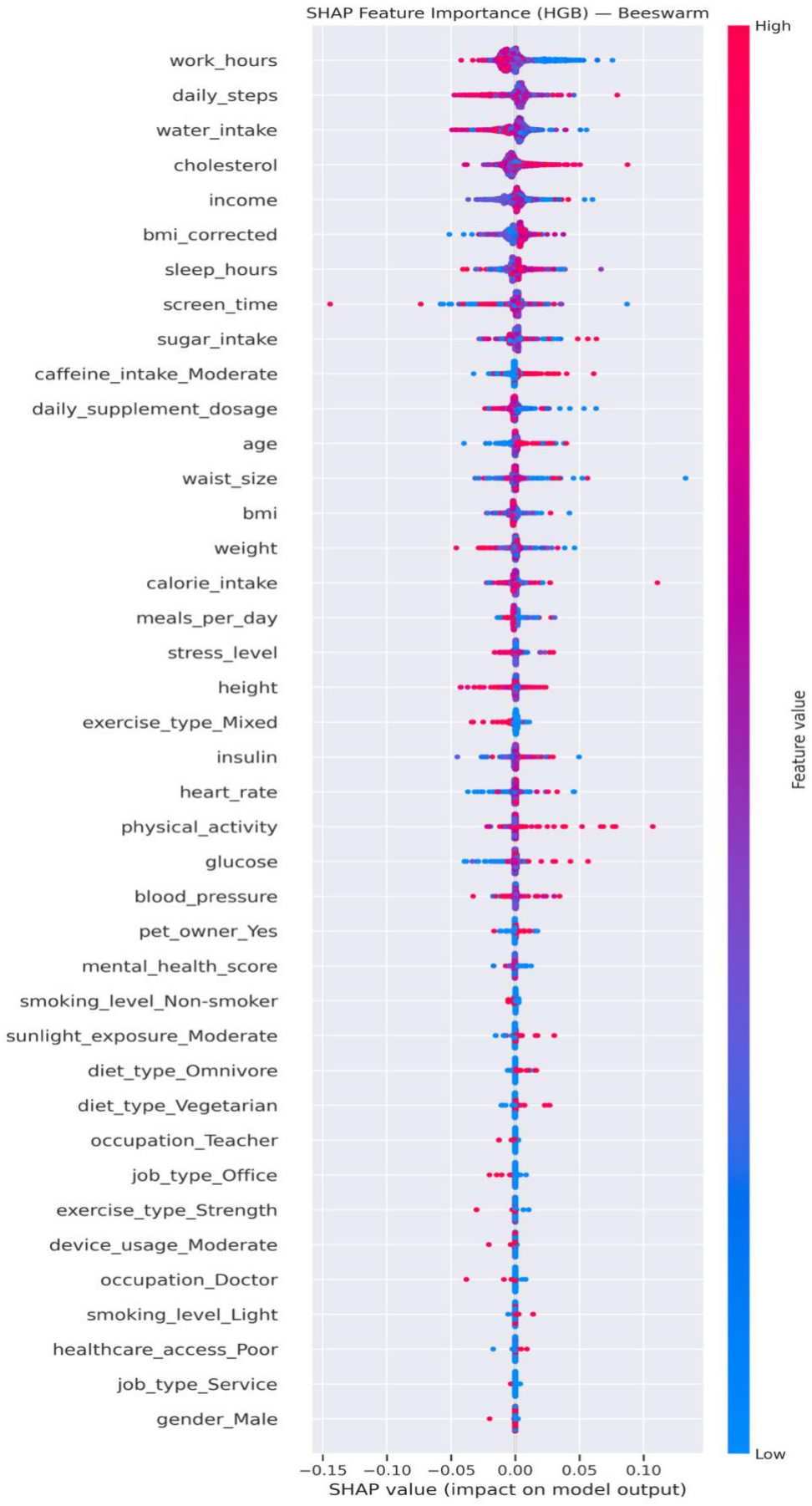
SHAP Feature Importance (HGB) — Beeswarm
work_hours
daily_steps
water_intake
cholesterol
income
bmi_corrected
sleep_hours
screen_time
sugar_intake
caffeine_intake_Moderate
daily_supplement_dosage
age
waist_size
bmi
weight
calorie_intake
meals_per_day
stress_level
height
exercise_type_Mixed
insulin
heart_rate
physical_activity
glucose
blood_pressure
pet_owner_Yes
mental_health_score
smoking_level_Non-smoker
sunlight_exposure_Moderate
diet_type_Omnivore
diet_type_Vegetarian
occupation_Teacher
job_type_Office
exercise_type_Strength
device_usage_Moderate
occupation_Doctor
smoking_level_Light
healthcare_access_Poor
job_type_Service
gender_Male
−0.15 −0.10 −0.05 0.00 0.05 0.10
SHAP value (impact on model output)
High
Feature value
Low

Figure 11: Precision-Recall Curve (Test Set) - RNN

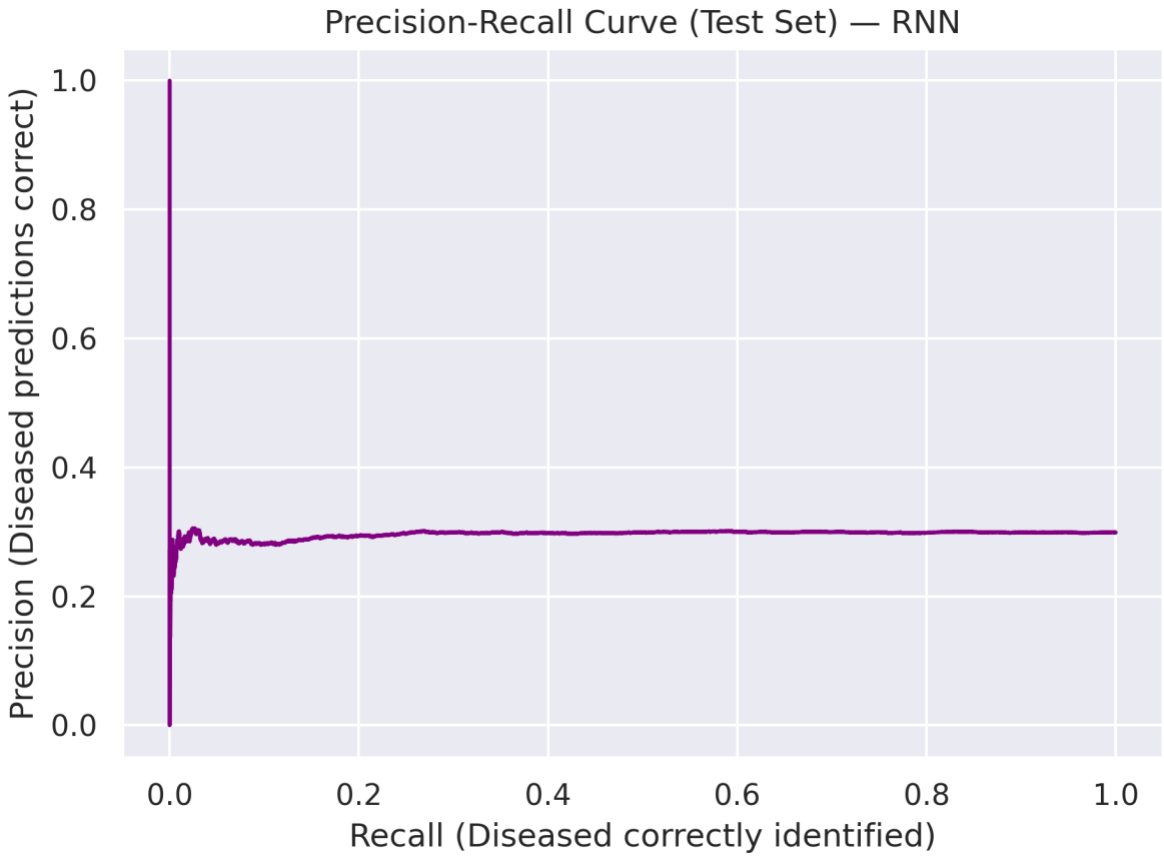

Figure 12: Precision-Recall Curve - HGB