A Prediction Framework for Lifestyle-Related Disease Prediction Using Healthcare Data

Lijuan Ren, Haiqing Zhang, Aicha Seklouli-Sekhari, Tao Wang, Abdelaziz Bouras

A Prediction Framework for Lifestyle-Related Disease Prediction Using Healthcare Data

LIJUAN REN, HAIQING ZHANG, AICHA SEKHARI SEKLOULI, TAO WANG, ABDELAZIZ BOURAS

***Abstract*—With the improvement of living standards and changes in work habits caused by industrialization, the prevalence of diseases linked to lifestyle is rising. In this context, the prevention of lifestyle-related diseases (LRDs) is extremely important. The majority of existing research exclusively concentrates on the prognosis of a particular LRD sickness, making it impossible for them to intelligently identify the important characteristics of the disease. Therefore, this study aims to propose a lifestyle-related disease prediction framework including three key components, called missing value module, a feature selection module, and a disease prediction module. The performance of the proposed framework is evaluated by using real medical data gathered during a hospital health check-up in Nanjing, China. The experiment shows that the proposed framework can automatically generate prediction ensemble models for specific LRDs diseases, and achieve good accurate performance.**

***Index Terms*—Lifestyle-related diseases, Prediction, Machine Learning, Missing values**

I. Introduction

Lifestyle-related diseases (LRDs) are illnesses that are significantly influenced by lifestyle factors, and changes in these factors can greatly improve disease prevention and treatment [1], [2]. As countries become more industrialized and affluent, which leads to bad lifestyles such as fast food, and sedentary, thereby the prevalence of LRDs is increasing. Most chronic diseases, including cardiovascular disease, metabolic syndrome, obesity, type 2 diabetes, and some cancers, are LRDs [2]. LRDs are currently the most common diseases in the world, and their death toll exceeds that of AIDS, malaria, and tuberculosis combined [3]. In the Republic of Ireland, over 40% of adults have at least one LRD, with high blood pressure and high cholesterol being the most prevalent [4]. In 2017, 17.8 million individuals globally died from cardiovascular disease (CVD), and the estimated number of tumor-related fatalities (mostly cancer) is 9.56 million [5]. The WHO predicts that by 2030, there will be 366 million individuals worldwide with diabetes, up from the current estimate of 175 million [6]. Despite the availability of numerous medications, LRDs remain uncontrolled due to safety concerns associated with these drugs [7]. Overall, the prevalence of LRDs represents a crisis in the global healthcare system.

Identify applicable funding agency here. If none, delete this.

Smoking, poor diet, excessive alcohol use, and a sedentary lifestyle are all clear contributors to various lifestyles related diseases [8], [9]. Many studies [10], [11] have shown that LRDs can be improved by healthy lifestyles. For example, Ford et al. [10] found that those who did not smoke, were not overweight, engaged in 3.5 hours of physical activity per week, and consumed a nutritious diet decreased their risks of myocardial infarction, stroke, cancer, and type 2 diabetes by 93%, 81%, 50%, and 36%, respectively, throughout the course of the 8-year trial. A study in Denmark aged 30 to 80 years showed that a change in physical activity level alone would result in an increase in life expectancy of between 2.8 and 7.8 years for men and between 4.6 and 7.3 years for women according to actual disease and death rates [11].

The idea of health is drastically altering, and the focus of healthcare is shifting from disease models to health models on a global scale [4]. Lifestyle-related diseases are multi-factorial illnesses that are influenced by environmental and genetic variables and brought on by the interaction of numerous risk factors [1]. These illnesses have sneaky onsets, a protracted incubation period, and a quick progression. Identifying and treating large numbers of patients in a timely manner is challenging. Additionally, as most lifestyle-related diseases still have unclear etiologies and pathogens and poor therapeutic outcomes, it is important from a practical standpoint to prevent the development of lifestyle-related diseases. Because identifying population risks prior to the onset of diseases can help people change their lifestyles as soon as possible, especially the life behaviors of high-risk groups, lowering the risk of disease [12]. The primary tool for assessing and preventing lifestyle-related diseases is the disease prediction model [13]. Disease prediction models specifically establish an intelligent model to predict the probability of a specific disease at a specific point in the future, classify high-risk groups in accordance with the probability cut-off point, and conducts behavior, diet, and other early interventions.

II. Research status

The original disease prediction model is a disease prediction model of coronary heart disease, which was established by the United States based on the Framingham cohort study [14], and other cardiovascular disease risk assessment models with various markers [15], [16]. The disease prediction models have gradually expanded from cardiovascular disease to include a variety of diseases [17]–[19]. Machine learning (ML) techniques, a subset of artificial intelligence techniques, employ computer systems to predict diseases using statistical models and algorithms, opening up a wide range of opportunities for illness prevention [12]. Researchers have utilized a number of ML algorithms to predict various diseases in the field of disease prediction. For instance, the use of ensemble techniques for the early diagnosis of coronary heart disease [20]; the use of support vector machines to detect pre-diabetes and diabetes [21]; the use of random forest algorithms to predict the risk of diabetes in the population examined physically [22]; To predict hypertension, a combination of subtype (the least absolute shrinkage and selection operator, LASSO) and support vector machine recursive feature elimination (SVMRFE) was used [23]. A new ensemble learning-based framework for the early detection of type 2 diabetes utilizing lifestyle markers was also developed [24].

However, existing prediction studies focus on single-disease prediction, with a few papers focusing on multiple-disease prediction. Yaganteeswarudu [25] proposed a system using the Flask API to predict multiple diseases including diabetes, diabetic retinopathy, heart disease, and breast cancer. This system uses different datasets to train different machine-learning models for different diseases. Rezaee M et al. [26] achieved consistent discrimination performance for multiple cardiovascular diseases and type-2 diabetes using prediction models derived from Cox proportional risk regression. These models contain multiple shared predictor variables and can be integrated into a single platform to enhance clinical stratification to influence health outcomes. Moreover, Rashid J et al. [27] proposed a new augmented artificial intelligence approach using artificial neural networks (ANN) and particle swarm optimization (PSO) to predict five prevalent chronic diseases including breast cancer, diabetes, heart disease, hepatitis, and kidney disease using five public datasets. Further, Gupta et al. employed a genetic algorithm based on recursive feature elimination and AdaBoost to predict two lifestyle diseases (heart disease and diabetes) using two public datasets with missing values.

Based on the above analysis, existing studies are unable to intelligently identify key features of diseases while building prediction models with different structures and robustness for different LRDs. Therefore, our objective is to design an intelligent risk prediction framework for LRDs that can smartly identify key features of different LRDs for dirty real medical data, and accurately predict the risk of LRDs.

III. The Overview of the Proposed Prediction Framework

A framework for LRDs prediction is proposed based on three key components, called missing value module, a feature selection module, and a disease prediction module. The method of combining deletion and imputation is chosen as the primary strategy for missing value processing for the significant number of missing values in the data set gathered from lifestyle-related diseases first. The feature selection module employs machine learning-based feature selection to discover key features for lifestyle-related diseases since different lifestyle-related diseases have distinct important features. In order to create a strong ensemble prediction model for lifestyle-related diseases and achieve a more accurate prediction of lifestyle-related diseases, the data processed by the missing value module and the feature selection module are used as the input of the prediction model. Figure 1 is a diagram of the proposed prediction framework for LRDs.

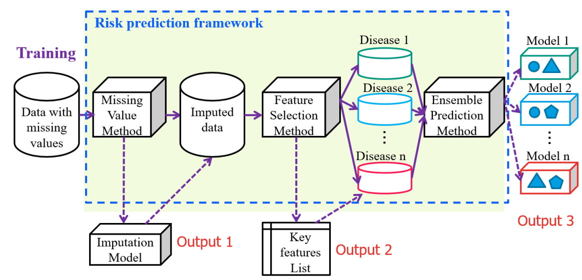

Fig. 1. LRDs prediction framework

In Fig.1, during the training process, the original dataset with different diseases is first feed into the missing value method to generate imputed data with higher quality. Then feature selection method is used to screen the resulting data to obtain sub-datasets with different features, which are passed into the final Ensemble Prediction Method to get target models adapting to different diseases. After feeding data, the framework has three outputs including an imputation model, a key feature list, and specific disease prediction models. Then, when new data comes, it can improve data quality by the trained imputation model, and then sub-datasets of key features corresponding to different disease models are input into target models to predict disease risk. After feeding data, the framework has three outputs, including an imputation model, a key features list, and some robust ensemble models for specific LRDs diseases. Then, when new data comes, it can improve data quality by the trained imputation model, and then sub-datasets of key features corresponding to different disease models are input into target models to predict disease risk. The application process of the proposed framework is shown in Fig.2.

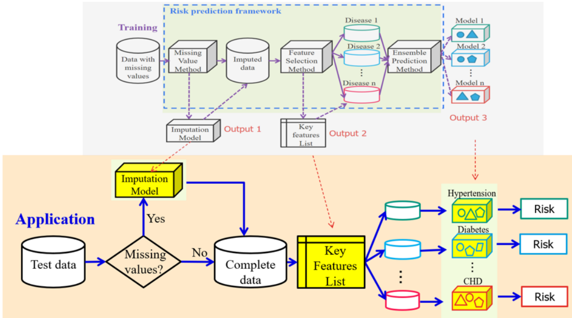

Fig. 2. The application process of the proposed framework

A. *Missing Value Module*

Some features or instances will have a disproportionate number of missing values for a variety of reasons, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [28]. The major features of lifestyle-related diseases are used in our study to build excellent predictive models, so when features or instances have a large number of missing values, this is difficult to apply in our study. Instead, we will prefer to use the deletion method rather than filling in a large number of estimates. We need to describe the criteria for deleting missing values, or the threshold for using it, in more detail. According to the 80% rule [29], which states that a substance should be removed if its non-missing portion is less than 80% of the sample size as a whole, the suggested prediction framework excludes features or instances whose missing rate is more than 80%.

There are still some missing values in the dataset even though some features and instances are compelled to be removed in accordance with the threshold setting of the missing rate. The reasons and ways of missing are typically dispersed among several features and instances, making it difficult to simply eliminate them using a deletion procedure. Therefore, to appropriately handle missing values, we shall employ more sophisticated techniques. In our previous study, we proposed a missing value imputation method, called SncALWRFI [30], that can be used with datasets that are imbalanced or mixed types. We employed this approach as our default missing value handling method in the missing value imputation step since features with characteristics of unbalanced and mixed types are common in datasets of lifestyle-related diseases. Similarly, we incorporate various well-known and excellent imputation methods for missing values, such as MissForest and KNNI, as alternatives or benchmarks to provide people with more options.

B. *Feature Selection Module*

As it can be challenging for people to distinguish between significant and superfluous features when gathering data, feature selection is an essential component of data reprocessing. Specifically, feature selection refers to choosing a task-related feature subset from the full set of features to reduce the amount of data that must be stored, shorten the time needed to train machine learning models, and enhance the predictive skills of machine learning models. Therefore, feature selection can assist in both the identification of essential features and the elimination of superfluous features. Data mining techniques based on machine learning techniques were used to select the primary characteristics of lifestyle-related diseases. The benefit of this approach is that the outcomes are generated by data analysis without the need for human interaction. This approach is appropriate for those without strong expertise in medicine and uses sophisticated algorithms to guide people in choosing essential factors. Our research belongs to the category of supervised learning because it focuses on the prediction of LRDs disease. We, therefore, concentrate on feature selection for supervised issues in this study. Three categories of feature selection techniques can be distinguished based on the form of the feature selection [31], and their advantages and disadvantages are shown in Table 1.

TABLE I
ADVANTAGES AND DISADVANTAGES OF THREE FEATURE SELECTION CATEGORIES

Category	**Advantage**	**Disadvantage**	**Example**
Filter	High efficiency	Ignore combination effect between features	Chi-Square
Wrapper	High accuracy	High complexity and over-fitting with small samples	Complete search
Embedded	Automatically selection	Determines loss function and parameters	Tree-based model

In Table 1, each method has its own advantages and disadvantages. The feature selection of the wrapper has high computation complexity, and the filtering mechanism ignores the connection between the feature and the target variable. As a result, the tree-based strategy in the embedding method is employed for feature selection in the proposed prediction framework. The proposed prediction framework uses the random forest importance approach as the main algorithm of the feature selection module because it can automatically identify features and is suited to mixed data types with high data dimensionality [32]. Specifically, the random forest feature importance evaluation calculates the mean value of each feature's contribution to each tree in the random forest. There are two techniques to obtain the final collection of key features after assessing the importance of each feature: 1) select Top-N features, 2) Select larger than the set threshold. Since the value of N is difficult to determine and in order to keep as many task-related features as possible, the feature selection module selects according to the important threshold of the feature.

C. *Disease Prediction Module*

As we previously mentioned, a variety of machine learning algorithms have been utilized by researchers to estimate the risk of various diseases in the field of disease risk prediction. In general, predictive models are divided into statistical-based and machine learning based and their advantages and disadvantages are shown in Table 2.

TABLE II
ADVANTAGES AND DISADVANTAGES OF PREDICTIVE MODELS

Category	Advantage	Disadvantage	Example
Traditional Statistical Model	Strong interpretability	Considering multiple assumptions; Poor modeling in complex data	Cox Regression
Machine Learning Model	High flexibility; High learning capability	High model complexity; Poor model interpretability	Support Vector Machines

Based on Table 2, to model complex data, our study adopted a machine-learning approach. On the other hand, it is challenging to employ a single model to generate more accurate forecasts and attain higher levels of performance due to the noise from attributes and classes. In machine learning, an ensemble is a sort of model that is built by merging the predictions of various individual models [33]. Typically, ensembles increase performance by reducing the mistakes created by each individual model that contributes to the ensemble. Therefore, to build a robust prediction model for LRDs, we will employ ensemble techniques to reduce the impact of noise. We employ a stacked ensemble method proposed in our previous study [34], a technique that can be used on datasets with diverse noise. This approach enables the data-driven selection of models to build integrated predictive models for different diseases.

IV. EXPERIMENT

A. Data Source

This study used real medical data gathered during a hospital health check-up in Nanjing, China. This dataset is from 2012 to 2022. All subjects in the study gave informed consent to the use of the data, and all sensitive information about the subjects was removed from the original dataset. In this real case study, hypertension, diabetes, and coronary heart disease are three common lifestyle-related diseases. First, we removed 23 records who were 20 years of age or younger. The remaining data comprised 32,784 instances and 65 attributes. Specifically, attributes include demographics (such as age, gender, and sex), urine tests (such as urine sugar, and urine occult blood), blood tests (such as glucose, and creatinine), and lifestyles (such as smoking, and drinking). Meanwhile, there are 18,936 males (57.75%) and 13,848 females (42.24%) in the dataset, with an age of 63.88±9.27.

For missing value analysis, 28% of the dataset's instances have less than 10% of their values missing, while 32.57 of them have missing values between 10% and 20%. Less than 0.02% of the instances lost more than 35% of the values at the same moment. Overall, no instance's portion of the dataset is missing by more than 50%, hence no instance is disregarded. 4 features' missing rate exceeds the 0.8 cutoff point, which means that 80% of their values are lost. The missing pattern in our case data is non-monotonic, which is also supported by the distribution plot of missing values. The findings of the missing value analysis show that, even after eliminating some features with 80% missing values, the data set still contains 8.84% missing values. Missing values are mainly distributed discretely in various measured features. It is not advised to delete the missing value model of the missing values in our data set directly since it is not missing completely at random (MCAR).

B. Missing Value Module

Firstly, we take hypertension as an example to analyze the effectiveness of the missing value module. The SncALWRFI imputation method was used to impute missing data based on the previous analysis. Pair deletion (PD), MEAN, KNNI [35], and MissForest [36] processing techniques were employed in comparison to examining the effects of the SncALWRFI imputation approach on the performance of lifestyle-related disease prediction. Because 80% of the instances contain missing values, the complete case analysis (CCA) approach is not employed because it is impossible to delete instances with missing values.

Additionally, to ensure fairness, default parameters are chosen for datasets processed by various missing value methods, along with Random Forest (RF), Light Gradient Boosting Machine (LGBM), and Logistic Regression Model(LRM) being used as predictive models for diseases connected to lifestyle. In detail, the data is split into two sets: a training data set, which comprises 70% of the data, and a testing data set, which contains 30% of the data. The training data set is used to create a missing value imputation model, and the test data set is used to assess the model's effectiveness. We compare performance using AUC as a performance indicator. The experiment was carried out 20 times, and Table 3 displays the average outcomes.

TABLE III
PREDICTION RESULTS OF DIFFERENT PROCESSING METHODS FOR MISSING VALUES

Methods	PD	MEAN	KNNI	MissForest	SncALWRFI
RF	75.02	80.07	81.10	82.72	**83.88**
LGBM	75.98	81.46	82.92	83.94	**84.83**
LRM	71.08	72.19	71.91	72.20	**73.31**

The maximum prediction result of 75.98 is obtained in the LGBM model, according to experimental results, while removing features with missing values yields the lowest prediction results. However, the SncALWRFI approach performs at its best, achieving an average ideal value of 84.83 in the LGBM model. Therefore, we employ SncALWRFI approach as the missing value processing method.

C. Feature Selection Module

Furthermore, the highly accurate and robust random forest-based feature selection (RF_FS) method is employed in the feature selection module. Specifically, the data without missing values preprocessed by the missing value module will be input, followed by the use of RF_FS to analyze the importance of features, and finally the selection of the data set containing only key features in accordance with the ranking of feature importance. A predictive model for LRDs was created using an experimental dataset. Initially, there were 65 features in our case, but since 4 of them (L_SQ, L_SA, L_DQ, and L_DA) were 80% absent from the dataset, they were excluded and the remaining 61 features were input into the feature selection module. When calculating feature importance, the result will be rounded to 3 decimal places. The top-N important features or all features with importance greater than 0 can be chosen once the calculation of feature importance is complete. In order to keep as many features as possible, the feature selection module selects according to the important threshold of the feature, that is, the features with importance of more than 0 are picked. Specifically, the RF_FS method selects 45, 38, and 43 important features for hypertension, diabetes, and coronary heart disease, respectively.

We take high blood pressure as an example as well, and the final experimental dataset will have 32,784 instances and 45 features. We use the same three prediction models and conduct 20 runs to confirm the impact of feature selection strategies on LRDs' prediction outcomes. The Table IV below displays the average AUC results obtained from 20 runs using various prediction models.

TABLE IV
PREDICTION RESULTS OF FEATURE SELECTION

Methods	RF	LGBM	LRM
Non - Feature Selection	83.88	84.83	73.31
Random Forest Feature Selection	**84.17**	**85.28**	**73.89**

The experimental results demonstrate that feature selection slightly increased the performance of the three prediction models, demonstrating that the feature selection method based on random forest can increase the accuracy of LRDs prediction after removing some features with low importance.

D. Disease Prediction Module

After analysis based on key features, the dataset with key features will be utilized to create a strong ensemble LRD predictive model. The final LRDs prediction model will be combined from candidate models including multilayer perceptron (MLP), K-Nearest Neighbors (KNN), Decision Tree (DT), support vector machine (SVM), Gaussian Bayese Network, Logistic LRM, Extreme Gradient Boosting (XGBoost), LightGBM, and RF. Three steps make up the model construction: ensemble model construction, hyperparameter optimization, and model evaluation. The disease prediction module will first automatically adjust the hyperparameters of each individual model in order to improve performance according to the parameter space in previous work [34]. Each model will receive the ideal set of parameters following the Bayesian optimization procedure. For hypertensive diseases, the disease prediction module will automatically generate a robust integrated prediction model and use the six-dimensional model capability map to automatically and visually evaluate the performance difference between the generated integrated model and the various sub-models that make up the model, as shown in Figure 3.

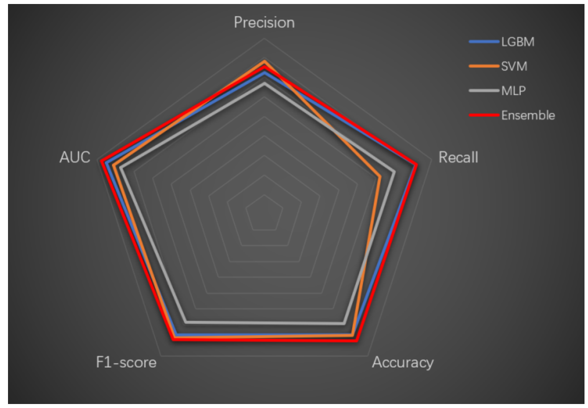

Fig. 3. Ensemble model evaluation

According to Figure 3, it can be seen that the constructed integrated model presents the best performance in the capability chart. And the value of the most important AUC index is 87.54, which shows that the model has a high discrimination ability and has practical application significance. Finally, to analyze the performance of the proposed disease prediction framework, we chose four advanced classifiers as the comparison methods. Meanwhile, we considered two missing value processing methods, including the deletion of features with missing values and the imputation method based on traditional random forest. The prediction results for the three diseases are shown in Table 4.

According to the table, we can observe that different missing value handling methods have little impact on the prediction performance, while the proposed framework can achieve the best prediction performance in the prediction of the three diseases. This demonstrates that adopting the proposed framework can intelligently identify key features of different LRDs against dirty real-world medical data, and automatically build robust prediction models for different lifestyle-related diseases to accurately predict the risk of LRDs.

V. CONCLUSION

Lifestyle Related Diseases (LRDs) refer to diseases whose pathophysiology is significantly affected by lifestyles. These diseases include diabetes, some malignant tumors, obesity, hypertension, coronary heart disease, other cardiovascular conditions, stroke, and other cerebrovascular conditions. Even with modern medication, such diseases pose a major threat to people's lives and health. The disease prediction model is the primary tool for evaluating and avoiding lifestyle-related diseases. Although a few works have addressed multi-disease forecasting, the majority of present forecasting studies concentrate on single-disease forecasting. As a result, we propose a new prediction framework for LRDs that consists of modules for disease prediction, feature selection, and missing values. Finally, we apply the proposed prediction methodology to a case from China. According to the experimental results,

TABLE V
THE PREDICTION RESULTS FOR THE THREE DISEASES

	RF		SVM		LRM		LGBM		Proposed
	PD	***MissForest***	***PD***	***MissForest***	***PD***	***MissForest***	***PD***	***MissForest***	***SncALWRFI***
Hypertension	75.02	82.72	82.24	84.21	71.08	72.20	75.98	83.94	**87.54**
Diabetes	77.16	79.30	79.88	81.33	69.35	71.62	72.32	80.61	**82.46**
CHD	79.59	80.26	81.63	81.69	79.47	79.75	81.01	82.22	**83.79**

the proposed prediction framework can automatically generate robust disease prediction ensemble models and perform well on LRDs prediction in healthcare data with missing values.

REFERENCES

[1] M Sagner, D Katz, G Egger, L Lianov, K Schulz, H, M Braman, B Behbod, E Phillips, W Dysinger, and D and Ornish. Lifestyle medicine potential for reversing a world of chronic disease epidemics: from cell to community. *International Journal of Clinical Practice*, pages 1289–1292, 2014.

[2] Byung-Il Yeh and In Deok Kong. The advent of lifestyle medicine. *Journal of lifestyle medicine*, pages 1–8, 2013.

[3] Abdallah S Daar, Peter A Singer, Stig K Persad, Deepa Leah Pramming, David R Matthews, Robert Beaglehole, Alan Bernstein, Leszek K Borysiewicz, Stephen Colagiuri, Nirmal Ganguly, Roger I Glass, Diane T Finegood, Jeffrey Koplan, Elizabeth G Nabel, George Sarna, Nizal Sarrafzadegan, Richard Smith, Derek Yach, and John Bell. Grand challenges in chronic non-communicable diseases. *Nature*, pages 494–496, 2007.

[4] G O'Donoghue, C Cunningham, F Murphy, C Woods, and J Aagaard-Hansen. Assessment and management of risk factors for the prevention of lifestyle-related disease: a cross-sectional survey of current activities, barriers and perceived training needs of primary care physiotherapists in the republic of ireland. *Physiotherapy*, 100(2):116–122, 2014.

[5] S Wilds, G Roglic, A Green, R Sicree, and H King. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care*, 27(5):1047–53, 2004.

[6] Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788, 2018.

[7] Shama Kakkar, Runjhun Tandon, and Nitin Tandon. The rising status of edible seeds in lifestyle related diseases: A review. *Food Chemistry*, page 134220, 2022.

[8] Ala Alwan et al. *Global status report on noncommunicable diseases 2010.* World Health Organization, 2011.

[9] Karen Morgan, Hannah Mcgee, Patrick Dicker, Ruairí Brugha, Mark Ward, Emer Shelley, Eric Van Lente, Janas Harrington, Margaret Barry, Ivan Perry, et al. Slán 2007: survey of lifestyle, attitudes and nutrition in ireland alcohol use in ireland: a profile of drinking patterns and alcohol-related harm from slán 2007. 2009.

[10] Earl S Ford, Manuela M Bergmann, Janine Kröger, Anja Schienkiewitz, Cornelia Weikert, and Heiner Boeing. Healthy living is the best revenge: findings from the european prospective investigation into cancer and nutrition-potsdam study. *Archives of internal medicine*, 169(15):1355–1362, 2009.

[11] Alvaro Sanchez, Paola Bully, Catalina Martinez, and Gonzalo Grandes. Effectiveness of physical activity promotion interventions in primary care: a review of reviews. *Preventive medicine*, 76:S56–S67, 2015.

[12] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

[13] Xiao-He Hou, Lei Feng, Can Zhang, Xi-Peng Cao, Lan Tan, and Jin-Tai Yu. Models for predicting risk of dementia: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(4):373–379, 2019.

[14] Jeanne Truett, Jerome Cornfield, and William Kannel. A multivariate analysis of the risk of coronary heart disease in framingham. *Journal of chronic diseases*, 20(7):511–524, 1967.

[15] Thomas J Wang, Philimon Gona, Martin G Larson, Geoffrey H Tofler, Daniel Levy, Christopher Newton-Cheh, Paul F Jacques, Nader Rifai, Jacob Selhub, Sander J Robins, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine*, 355(25):2631–2639, 2006.

[16] Paul M Ridker, Julie E Buring, Nader Rifai, and Nancy R Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *Jama*, 297(6):611–619, 2007.

[17] Philip A Wolf, Ralph B D'Agostino, Albert J Belanger, and William B Kannel. Probability of stroke: a risk profile from the framingham study. *Stroke*, 22(3):312–318, 1991.

[18] PM Clarke, AM Gray, A Briggs, AJ Farmer, P Fenn, RJ Stevens, DR Matthews, IM Stratton, and RR Holman. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the united kingdom prospective diabetes study (ukpds) outcomes model (ukpds no. 68). *Diabetologia*, 47(10):1747–1759, 2004.

[19] Margaret R Spitz, Waun Ki Hong, Christopher I Amos, Xifeng Wu, Matthew B Schabath, Qiong Dong, Sanjay Shete, and Carol J Etzel. A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, 99(9):715–726, 2007.

[20] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J Khoury. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, 10(1):1–7, 2010.

[21] Vardhan Shorewala. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26:100655, 2021.

[22] Zhanlin ZHANG, Yong SUN, Xiaoqing TUO, et al. Predictive value of random forest algorithms for diabetic risk in people underwent physical examination. *Chinese General Practice*, 22(9):1021, 2019.

[23] Md Merajul Islam, Md Jahanur Rahman, Dulal Chandra Roy, Most Tawabunnahar, Rubaiyat Jahan, NAM Faisal Ahmed, and Md Maniruzzaman. Machine learning algorithm for characterizing risks of hypertension, at an early stage in bangladesh. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(3):877–884, 2021.

[24] Shahid Mohammad Ganie and Majid Bashir Malik. An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, 2:100092, 2022.

[25] Akkem Yaganteeswarudu. Multi disease prediction model by using machine learning and flask api. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1242–1246. IEEE, 2020.

[26] Mehrdad Rezaee, Igor Putrenko, Arsia Takeh, Andrea Ganna, and Erik Ingelsson. Development and validation of risk prediction models for multiple cardiovascular diseases and type 2 diabetes. *PLoS one*, 15(7):e0235758, 2020.

[27] Junaid Rashid, Saba Batool, Jungeun Kim, Muhammad Wasif Nisar, Amir Hussain, Sapna Juneja, and Riti Kushwaha. An augmented artificial intelligence approach for chronic diseases prediction. *Frontiers in Public Health*, 10:559, 2022.

[28] Roderick J.A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics, 2014.

[29] Sabina Bijlsma, Ivana Bobeldijk, Elwin R Verheij, Raymond Ramaker, Sunil Kochhar, Ian A Macdonald, Ben Van Ommen, and Age K Smilde. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical chemistry*, 78(2):567–574, 2006.

[30] Lijuan Ren, Aicha Sekhari Seklouli, Haiqing Zhang, Tao Wang, and Abdelaziz Bouras. An adaptive laplacian weight random forest imputation for imbalance and mixed-type data. *Information Systems*, 111:102122, 2023.

[31] Khaled Mohamad Almustafa. Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, 24:100631, 2021.

[32] Mehrdad Rostami and Mourad Oussalah. A novel explainable covid-19 diagnosis method by integration of feature selection with random forest. *Informatics in Medicine Unlocked*, 30:100941, 2022.

[33] Fei Li, Li Zhang, Bin Chen, Dianzhu Gao, Yijun Cheng, Xiaoyong Zhang, Yingze Yang, Kai Gao, and Zhiwu Huang. An optimal stacking ensemble for remaining useful life estimation of systems under multi-operating conditions. *IEEE Access*, 8:31854–31868, 2020.

[34] Lijuan Ren, Haiqing Zhang, Aicha Sekhari Seklouli, Tao Wang, and Abdelaziz Bouras. Stacking-based multi-objective ensemble framework for prediction of hypertension. *Expert Systems with Applications*, 215:119351, 2023.

[35] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[36] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.