DATA-DRIVEN MODELING AND ANALYSIS FOR CARDIOVASCULAR DISEASE RISK PREDICTION AND REDUCTION

By

AYSE DOGAN

Bachelor of Science in Industrial Engineering

with double major

Bachelor of Arts in Economics

Antalya Bilim University

Antalya, Turkey

2019

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2021

DATA-DRIVEN MODELING AND ANALYSIS FOR CARDIOVASCULAR DISEASE RISK PREDICTION AND REDUCTION

Thesis Approved:

Dr. Chenang Liu

Thesis Adviser

Dr. Manjunath Kamath

Dr. Bing Yao

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Chenang Liu, for his guidance and mentoring throughout my research. I really appreciate his concern about the quality of my work and for his rigorous research attitude and concern about my academic and professional goals. I would also like to thank my committee member, Dr. Bing Yao, for her kind advices. I would also like to thank Dr. Manjunath Kamath for his guidance as my committee member. I would also like to thank Dr. Ying Lin, Mr. Yuxuan Li, Mr. Zhangyue Shi and Mr. Chiwetalu Odo, and Ms. Sule Nur Kutlu for their assistance and comments during my research. I would also like to thank the entire Industrial Engineering and Management department staff for their administrative support. Last, but not least, I would like to thank my family for their support, patience, and confidence.

Name: AYSE DOGAN

Date of Degree: MAY 2021

Title of Study: DATA-DRIVEN MODELING AND ANALYSIS FOR CARDIOVASCULAR DISEASE RISK PREDICTION AND REDUCTION

Major Field: INDUSTRIAL ENGINEERING & MANAGEMENT

Abstract: In recent decades, cardiovascular disease (CVD) has become the leading cause of death in most countries of the world. Since many types of CVD could be preventable by modifying lifestyle behaviors, the objective of this study is to develop an effective personalized lifestyle recommendation approach for reducing the risk of common types of CVD. However, in practice, the underlying causal relationship between the risk factors (e.g., lifestyles, blood pressure, etc.) and disease onset is highly complex. Furthermore, it is also challenging to identify the most effective modification for different individuals by considering both individual's preferences and the uncertainties in disease progression. Therefore, to address these challenges, this study developed a novel data-driven approach for personalized lifestyle behaviors recommendation based on machine learning and utility function model. The contribution of this work can be summarized into three aspects: (1) a classification-based prediction model is implemented to accurately predict the CVD risk based on the condition of risk factors; (2) a GAN-based approach is developed to capture the relationship between risk factors and generate feasible healthier lifestyle modifications; and (3) a novel personalized evaluation model incorporating utility function is proposed to identify the optimal modification with the consideration of individual's cost of change and disease progression uncertainty. The effectiveness of the proposed method is validated through an open-access CVD dataset. The results demonstrate that the personalized lifestyle recommended by the proposed methodology can significantly reduce the potential CVD risk. Thus, it is very promising to be further applied to real-world cases for CVD prevention.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Cardiovascular disease (CVD) is a class of common chronic diseases that involve the heart or blood vessels, such as myocardial infarction (MI), stroke, coronary heart disease (CHD), etc. It has been one of the most frequent causes of death in the United States for a long time [1], e.g., sudden cardiac death. Although CVD poses a great challenge for human health, existing studies, such as Ref. [2], have shown that most types of CVD could be potentially well-prevented if the related risk factors can be controlled appropriately [3], particularly, for the lifestyle behaviors, such as tobacco, diet, exercise.

Existing guidelines for CVD prevention were developed for the whole population, which may not be the optimal option for each individual. To help vulnerable individuals prevent CVD more effectively, a new methodology that can provide personalized suggestions for lifestyle behavior modification is critically needed. In practice, the underlying biomedical mechanism of the causal relationship between lifestyle and CVD risk is very difficult to be quantitatively described. However, investigations for the individual-based healthcare database by applying advanced data analytics offer a promising research opportunity. For example, the Atherosclerosis Risk in Communities (ARIC) study [4] integrates the survey data, clinical test, medical diagnosis, and population surveillance data for the investigation about CVD occurrence and patients' lifestyle.

Therefore, the objective of this study is to develop an effective data-driven personalized lifestyle recommendation approach for reducing the risk of CVD. However, there are three practical major challenges: (1) the CVD risk for each individual is unknown; (2) the relationship between risk factors is highly complex; and (3) it lacks an effective personalized evaluation framework for the selection of optimal lifestyle recommendation.

To address these challenges, this study developed a new data-driven personalized recommendation approach by the generative adversarial network (GAN) [5] and a novel personal-regularized utility function model for CVD prevention. Recently GAN has become one of the most popular machine learning models for learning the complex distribution of multivariate data. In the proposed method, GAN is able to generate alternative lifestyle behaviors by considering the joint distribution of the risk factors. Meanwhile, it is also applied to synthesize the important physical examination indices after the potential modification, which then enables the CVD risk prediction for the alternative lifestyle behaviors. Furthermore, based on the expected utility theory, which is a popular decision-making approach, a novel personal-regularized utility function model is proposed to select the optimal lifestyle modification by jointly considering an individual's CVD risk, cost of change, and disease progression uncertainty.

The rest of this study is organized as follows. A brief review of the research background and related work is provided in Sec. 2. The data is described in Section 3. The proposed research methodology is presented in detail in Sec. 4. Subsequently, the results and discussions are demonstrated in Sec. 5. Finally, conclusions and possible future work are presented in Sec. 6.

CHAPTER II

REVIEW OF RELATED WORK

This study is focused on lifestyle recommendation system development for reducing the risk of CVD. Thus, this section first briefly reviewed the related studies on the effect of lifestyle behavior on cardiovascular disease (Sec. 2.1) then it is followed by a review of the existing data-driven methodologies for a recommendation (Sec. 2.2).

2.1 Effect of lifestyle behavior on CVDs

To reduce the cause of heart failure, the effects of lifestyle behaviors on CVD risk have become one of the most critical areas in heart disease research. Several popular long-term cohort studies regarding heart disease, including the Framingham heart study (FHS) [8] and Atherosclerosis Risk in Communities (ARIC) Study, provided great data resources and research opportunities to investigate the CVD risk factors. For example, through FHS, Vasan *et al.* [9] explored the correlation between antecedent blood pressure and the risk of CVD, and Wilson *et al.* [10] investigated the effect of blood pressure and cholesterol on CHD risk together. Regarding the long-term risk prediction, Pencina *et al.* [11] developed a statistical model using the FHS data to predict the CVD risks in 10 years and 30 years. Wickramasinghe *et al.* [12] further considered the fitness level for 30-year risk cardiovascular mortality prediction. Lloyd-Jones [13] also provided a review of the history and principles of CVD risk prediction.

Furthermore, considering the effects of lifestyle behaviors, E Millen *et al.* [14] examined the relationship between dietary patterns, food frequency, and atherosclerotic disease by using the FHS. Recently, more studies are conducted based on the ARIC data, for example, Chi *et al.* [5], Mansoor *et al.* [15], and Chambless *et al.* [16] investigated the effect of different lifestyles on CVD risks. In these studies, different machine learning and statistical modeling methods are applied, including k nearest neighbors, Kaplan-Meier analysis, and Cox regression model. Although the above-mentioned studies made great progress in the analysis of CVD risk in recent decades, it still lacks effective methods to provide individualized guidelines, particularly, lifestyle modification plans, for CVD prevention.

2.2 Data-driven recommendation approaches

Besides the risk prediction and assessment, it is also critical to provide personalized recommendations for health improvement, which has been investigated in several existing studies, including Enwald *et al.* [17], Skinner *et al.* [18], Kreuter *et al.* [19], Brug *et al.* [20], and Oenema *et al.* [21]. For example, Enwald *et al.* [17] presented that nutrition and physical exercises are supportive tailoring interventions, which have a positive health effect since the individuals could have a more active role in their health. However, these studies assumed that all the features can be directly affected by human intervention. Besides, the correlations between different features are also not considered.

Recently, data-driven recommendation approaches have become more and more popular in a large variety of areas, and they also demonstrate great potential for recommending alternative lifestyles to reduce the risk of diagnosing with diseases for individuals. For example, using the data provided by cardiologist, Jabeen *et al.* [22] developed an approach for CVD treatment suggestion based on a perspective of the Internet of Things. Another perspective, using the ARIC database, Chi *et al.*

[5] developed an expert system for lifestyle recommendations to reduce the CVD risk using the k-NN algorithm. However, k-NN is lazy and not able to measure the effects of features on the outcomes independently from each other. Nam *et al.* [23] also developed an expert system for monitoring physical activity and recommending lifestyle interventions in obesity. The developed system is able to output a diet and exercise program for treating obesity.

Among various data-driven recommendation methodologies, as an emerging technique, the generative adversarial network (GAN) recently becomes one of the most famous choices. The GAN-based recommendation algorithms have been developed and applied in many fields with its different applications with different recommendation types, such as Kang *et al.* [24], He *et al.* [25], and Li *et al.* [26]. Furthermore, recent studies also try to integrate GAN with other powerful machine learning techniques to further improve the recommendation accuracy and effectiveness. For example, reinforcement learning (RL)-based recommendation [27] is one of them to develop imitated user behaviors dynamically and learn the reward function for it. The RL policy can provide a better long-term reward for the users. Besides, the widely applied Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) are also successfully incorporated with GAN, namely, RecGAN [28] and DRCGR [29], respectively. GAN has been successfully applied for these various recommendation systems, as a very powerful technique, its applications in lifestyle recommendation for reducing disease risk, particularly, CVD risk, is still very limited.

CHAPTER III

DATA DESCRIPTION

The data used in this study is extracted from the Cohort Component of the Atherosclerosis Risk in Communities (ARIC) [4], which was collected from four US communities: (1) Forsyth Country, NC; (2) Jackson, MS; (3) Suburban Minneapolis, MN; and (4) Washington Country, MD [30] as it is shown in Figure 1.

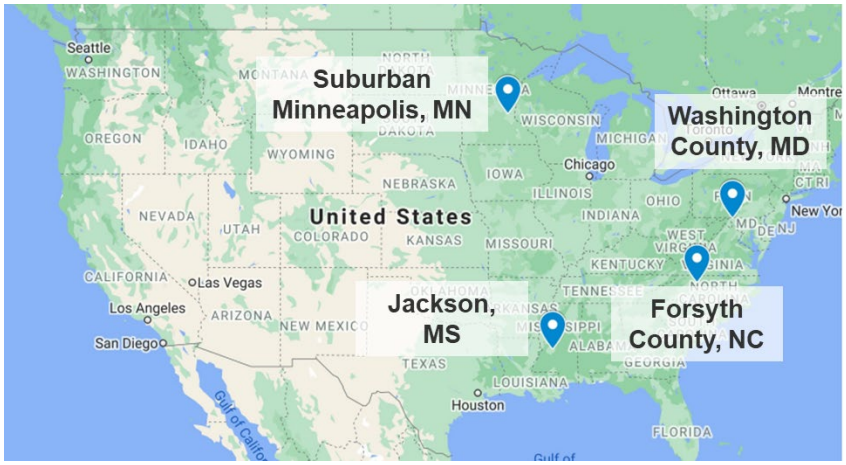

Figure 1: Regional display of the ARIC participants

In each community, approximately 4,000 individuals aged 45-64 were randomly selected and recruited to collect their medical, social, and demographic data. In this study, the data collected from the participants' first visit during 1987-1989 (i.e., visit 1) are selected for the methodology development and validation.

Meanwhile, participants who are already diagnosed with any type of heart failure or CVD before the first visit are removed. Therefore, in total 13,654 participants are included. Subsequently, the risk factors that may impact the CVD risk are identified and selected from the dataset based on the literature [1, 13, 16, 31]. The selected risk factors consist of three categories that are obtained from health records: directly changeable variables, indirectly changeable variables, and unchangeable variables as it is shown in Figure 2.

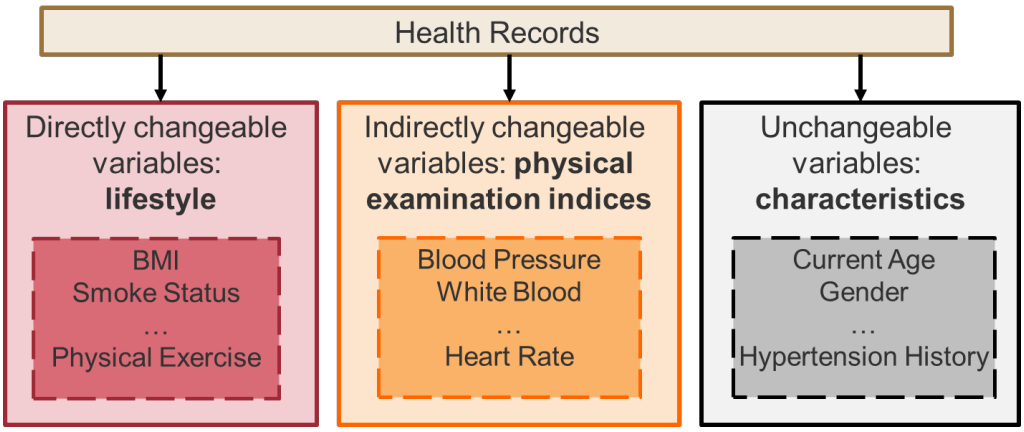

Figure 2: Data Architecture

The directly changeable variables describe each participant's lifestyle behaviors, including diet, exercise, cigarette, alcohol, etc. In general, these variables can be directly changed by individuals' efforts based on the recommendations from physicians. As mentioned in Sec. 1, this study aims to identify the optimal modifications of each individual's lifestyle behaviors, which is essential to appropriately adjust the directly changeable variables. In this study, 11 directly changeable variables are selected from the ARIC database based on the existing literature [13, 31] that can be seen in Table 1.

Table 1: The list of the directly changeable features

Number	Type	Name	Definition
1	NUM	ETHANL03	Alcohol intake (g) per day
2	Nominal	SMKSTA	Smoking status [0, 1]
3	NUM	WORK_I02, SPRT_I02, LISR_I01 (EXE01)	Total activity hours per week
4	NUM	CARB	Carbohydrate (g)
5	NUM	PROT	Protein (g)
6	NUM	SFAT	Saturated fatty acid (g) per day
7	NUM	TFAT	Total fat (g) per day
8	NUM	BMI01	Body mass index
9	NUM	CHOL	Dietary cholesterol (mg)
10	NUM	DFIB	Dietary fiber (g)
11	NUM	TCAL	Total Energy Intake (mg/dL)

These variables are denoted as,

$$\mathbf{X}_D = (\mathbf{x}_D^1, \mathbf{x}_D^2, \cdots, \mathbf{x}_D^{11}) \quad (1)$$

where $\mathbf{x}_D^i$ represents each directly changeable variable in this study. Considering the modification applicability in practice, the continuous variables in $\mathbf{X}_D$ are discretized to five intervals based on quantiles, while the categorical variables remain the same. For example, $\mathbf{x}_D^1 = \text{BMI}$ is split into 5 interval based on these borders; [14.201, 23.162, 25.533, 27.862,31.214, 65.91].

Then for the indirectly changeable variables, denoted as $\mathbf{X}_I$, most of them are the common physical examination indices related to the health of participants, such as systolic and diastolic blood pressures, apolipoprotein, creatinine, etc., shown in Table 2,

Table 2: The list of indirectly changeable features

Number	Type	Variable	Definition
1	NUM	HDLSIU02	HDL cholesterol in mg/dl
2	NUM	LDLSIU02	LDL cholesterol in mg/dl
3	NUM	TCHSIU01	Total cholesterol in mmol/L
4	NUM	TRGSIU01	Total triglycerides in mmol/L
5	NUM	SBPA21	2nd and 3rd Systolic blood pressure average
6	NUM	SBPA22	2nd and 3rd Diastolic blood pressure, blood pressure average
7	NUM	ANTA07A	Waist girth to nearest CM
8	NUM	ANTA07B	Hip girth to nearest CM
9	NUM	ECGMA31	Heart rate
10	NUM	HMTA03	White blood count
11	NUM	APASIU01	Apolipoprotein AI (MG-DL)
12	NUM	APBSIU01	Apolipoprotein B (MG-DL)
13	NUM	LIPA08	APOLP (A) DATA (UG-ML)
14	NUM	CHMA09	Creatinine (MG-DL)

In this study, the indirectly changeable variables are denoted as,

$$\mathbf{X}_I = (\mathbf{x}_I^1, \mathbf{x}_I^2, \cdots, \mathbf{x}_I^{14}) \tag{2}$$

where $\mathbf{x}_I^i$ shows each of the indirectly changeable variables. Notably, these variables may also be affected by the changes of directly changeable features. For instance, there is a strong causal relationship between blood pressure and some directly changeable features such as dietary and activity hours [32]. Therefore, participants can indirectly reduce or increase these variables by changing the directly changeable variables. Lastly, some variables cannot be changed since they are the characteristics of the individuals, e.g., people cannot change their ages at the time, if they have been diagnosed with Hypertension or Diabetes this will remain in their health history. Thus, these variables are named unchangeable features, $\mathbf{X}_U$. The detailes are shown in Table 3.

Table 3: The list of unchangeable features

Number	Type	Variable	Definition
1	NUM	CIGTYR01	Cigarette years of smoking
2	Nominal	ELEVEL01	Education level [(1) grade school or 0 years education, (2) high school, but no degree, (3) high school graduate (4), vocational school, (5) college (6) graduate school or professional school]
3	Nominal	GENDER	Sex
4	Nominal	RACEGRP	Race
5	NUM	V1AGE01	Age
6	Nominal	DIABTS02	Diabetes [0, 1]
7	Nominal	HYPERT04	Hypertension [0, 1]
8	Nominal	HYPTMDCODE01	Hypertension medication in the past 2 weeks
9	Nominal	CHOLMDCODE01	Cholesterol-lowering medication use [0, 1]
10	NUM	ANTA01	Standing height to nearest CM
11	Nominal	CIGT01	Smoking status history
12	NUM	HYPTMD01	HYPERTENSION LOWERING MED. USE
13	NUM	ANTICOAGCODE01	Used Anticoagulates (at Visit 1) last 2 weeks (0=No, 1=Yes) based on 2004 Med Code
14	NUM	ASPIRINCODE01	Used Aspirin-containing analgesics (at Visit 1) in last 2 weeks (0=No, 1=Yes), based on 2004 Med Code
15	NUM	STATINCODE01	Used Statin (at Visit 1) last 2weeks (0=No, 1=Yes) based on 2004 Med Code

Unchangeable variables are also denoted in a similar way,

$$\mathbf{X}_U = (\mathbf{x}_U^1, \mathbf{x}_U^2, \cdots, \mathbf{x}_U^{15}) \tag{3}$$

where $\mathbf{x}_U^1$. Since the unchangeable variables are representing a certainty they are effective on $\mathbf{X}_U$ variables. Furthermore, four 10-year CVD outcomes (y_i, $i = 1,2,3,4$) are defined using the Community Surveillance Component in ARIC, which indicate the occurrence of Fatal CHD (y_1), MI (y_2), and Stroke (y_3) (that indicates at least one of the ischemic, incident, or hemorrhagic types of Stoke), as well as any type of them (y_4) within the 10 years follow-up.

CHAPTER IV

PROPOSED RESEARCH METHODOLOGY

As shown in Figure 3, the overall framework of the proposed research methodology consists of three steps: (1) appropriate features are selected from the ARIC database, and the outcome variables are defined as well (Sec. 3); (2) a classification-based predictive model is developed for individual CVD risk assessment (Sec. 4.1); and (3) a personalized alternative lifestyle recommendation approach by incorporating GAN and a regularized utility function model is proposed to reduce the individual's CVD risk. (Sec. 4.2). Additionally, the approach to validate the effectiveness of the proposed methodology is also discussed in Sec. 4.4.

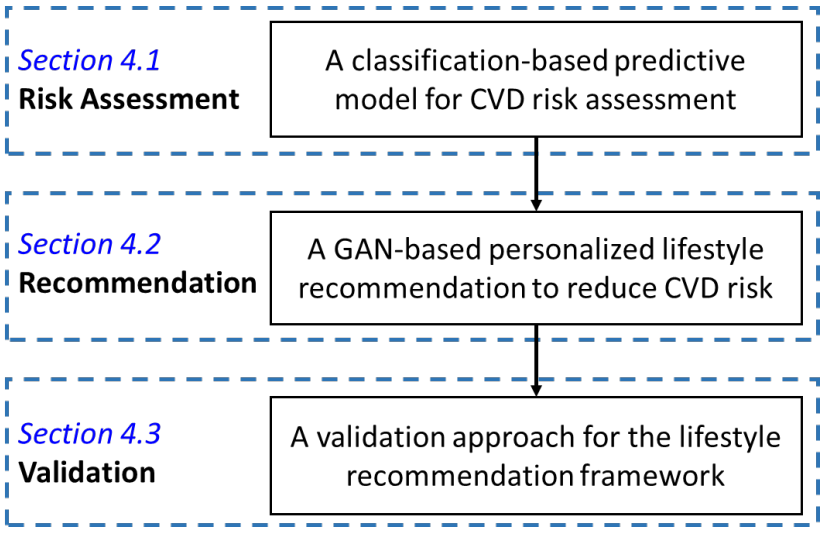

***Figure 3**: The overall framework of the proposed research methodology.*

4.1 Classification model for CVD risk estimation

To identify the high-risk patients and evaluate the effectiveness of the suggested lifestyle modifications, the CVD risk of each participant needs to be accurately predicted using the selected risk-related variables ($\mathbf{X}_I, \mathbf{X}_U, \mathbf{X}_D$), i.e., to develop a predictive model, as formulated in Eq. (4),

$$r = P(y_i = 1) = f(\mathbf{X}) = f(\mathbf{X}_I, \mathbf{X}_U, \mathbf{X}_D) \tag{4}$$

where f is an unknown nonlinear function to predict the CVD risk. To train an accurate f using the collected data after imputing the missing values, supervised machine learning techniques, i.e., classification algorithms are applied in this study.

4.1.1 Missing Value Imputation

In practice, due to several practical reasons, for example, rejected responses to some survey questions or some unincluded test results, missing values are existing in the collected dataset. Consequently, to ensure model accuracy, it is necessary to impute the missing values before training f. In this study, a widely applied data-driven imputation approach, namely, MissForest [33] and k nearest neighbor (k-NN) [34] methods are applied and compared.

The MissForest is an imputation method whose working strategy operates based on the random forest algorithm. It is good to cope up with nonlinear relations and complex interactions that can be very useful for a healthcare data imputation without using any predetermined parameter.

After that, k-NN imputation estimates the missing value by k assigned samples (nearest neighbors) that are similar or close in the dataset, and then it can be imputed using the mean value of these k

neighbors in the dataset [34]. In practice, the value of k plays a significant role since a smaller k may cause underfitting while a larger one may result in overfitting.

To select the best imputation method for missing values, MissForest and k-NN imputation with many different k values are applied to the data. To select an imputation method classification outcomes have been compared. From the classification outcomes, the data that is imputed with the k-NN imputation method gives the best performance. Thus, k-NN imputation is selected to impute the missing values in this study.

4.1.2 Risk predictions with classification algorithms

After data imputation is performed the data is ready for the classification applications. From the widely-applied classification algorithms, Random Forest [35] is a very popular supervised classification method that consists of multiple decision trees. These multiple trees aim to reduce the possible overfitting that can be obtained from a single decision tree. Since it is supervised it needs to be trained at some rate of the dataset. And rest of the data set can be used while testing the performance of the algorithm. Let's say there are n features in the dataset. By the bootstrapping method, the random multiple samples are created from the main training dataset for each tree which is the classifier. After that, each of the trees decides a class for the output. At the end of the algorithm, the random forest returns the output based on the ensemble of the n classifiers, as shown below in Figure 4.

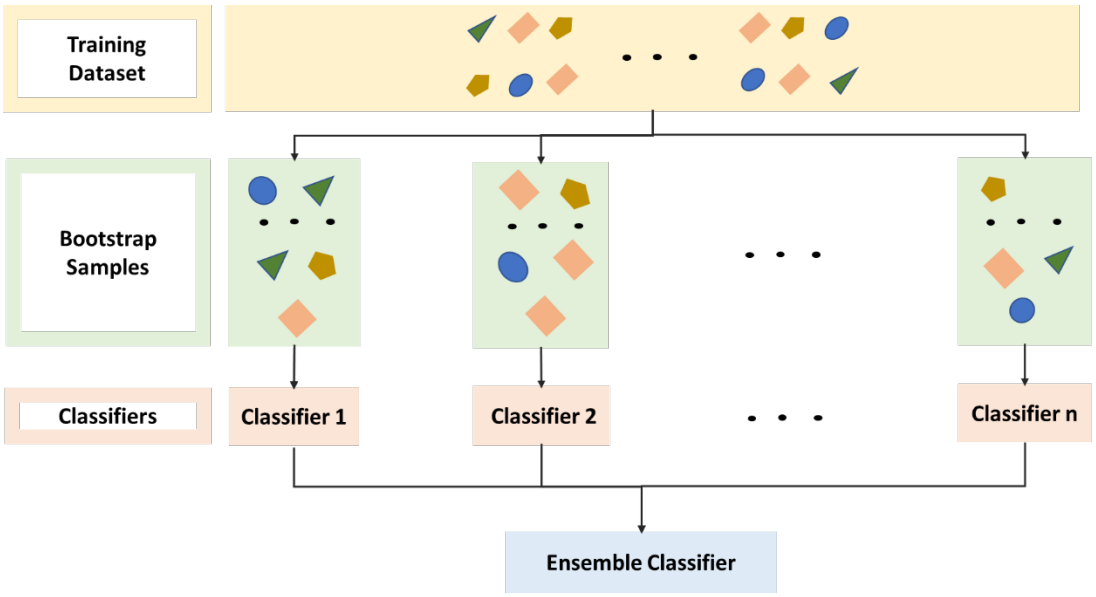

Figure 4: *Random Forest (RF) Classifier working logic*

After training, to test the algorithm, the test dataset can be input for the ensemble classifier which can return the predictions or the classifications. Other than the random forest, the support vector machine (SVM) [36], and k-NN classification [37], are also applied to fit f after missing value imputation.

For model comparison and selection of the classifiers, the receiver operating characteristic (ROC) curve [38] and its AUC (area under the curve) score [38], are utilized. ROC curve is a curve of probability that displays the true positive rate vs. the false positive rate. The AUC is a metric that cites the degree or measure of separability of the ROC curve. Based on the AUC score, it can be seen how much the model is capable of distinguishing the classes between each other. A higher AUC score shows a better model to predict the outcome classes. It is also a safe way to make a model comparison. Through comparison (see more details in Section 5.1), the random forest is selected to fit a CVD risk prediction model $\hat{f}$ that is approximate to f.

4.2 Personalized lifestyle recommendation using GAN and regularized utility function

Based on the developed CVD risk prediction model introduced in Sec. 4.1, the expected risk of the potential alternative lifestyle behaviors could be evaluated for the patients. However, there are still two major challenges to identify the optimal alternative lifestyle for each patient: (1) the complex interactions between $\mathbf{X}_I$, $\mathbf{X}_U$, and $\mathbf{X}_D$ pose great uncertainty in disease progression, and (2) the willingness and cost to change lifestyles of the patients should be considered as well. To address these two challenges, a novel personalized lifestyle recommendation approach by incorporating the powerful GAN and a new regularized expected utility function model is proposed in this section.

4.2.1 Generative adversarial networks (GAN)

As a popular emerging machine learning model, GAN has demonstrated its strong capability of capturing the complex underlying relationship between variables. The main idea of GAN [5] is to train two networks, namely, generator G and discriminator D, with a minimax game for D and G demonstrated in Eq. (5),

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{data}}(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))] \quad (5)$$

where the generator G generates artificial samples $\mathbf{z}$ and the discriminator D aims to label the input samples $\mathbf{x}$. G is trying to trick the D by its artificial samples and make D be not able to distinguish whether it is artificial or actual. In another word, G is aiming to produce artificial data that is so similar to original ones.

In this study, an extension of GAN, namely, the generative adversarial imputation networks (GAIN) [39], is applied to implement the following two tasks: (1) generate applicable alternative

lifestyle behaviors for the risky individuals, i.e., generate alternative $\mathbf{X}_D$ by a given $\mathbf{X}_U$; and (2) generate feasible indirectly changeable variables associated with the generated alternative lifestyle behaviors for the risky individuals, i.e., generate $\mathbf{X}_I$ by a given $(\mathbf{X}_D, \mathbf{X}_U)$.

The GAIN algorithm is one of the most popular imputation algorithm these days [39]. In GAIN, G observes the actual data and generates the needed data according to the observed data. Afterward, D takes the completed data from G, and tries to discriminate which part of data is generated and which part is actual. Meanwhile, the random input Z provides randomness to the G to not make the imputations the same every time. Apart from that, the hint matrix H makes the connection between D and G. The aim of the hint is making D be able to distinguish the imputed variables of G from the present ones. In this way, D forces the G predict and impute better. A basic explanation of the model is shown in Figure 5 below.

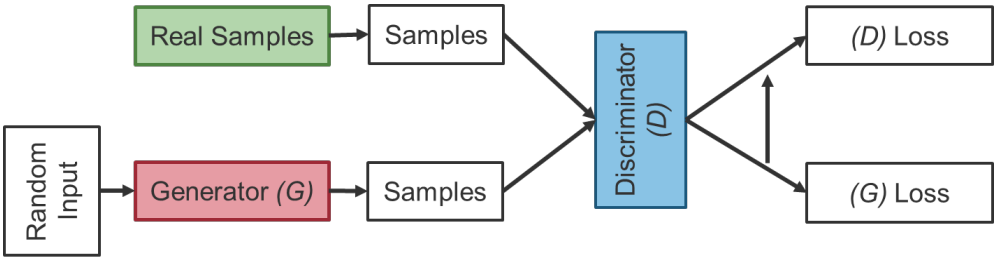

Figure 5: *The summary schema of the GAN method.*

Using the collected training data, GAIN will learn the underlying relationships between variables, so that it will gain the capability to accomplish the above-mentioned two tasks. More details of GAIN could be found in Ref. [39].

Based on the trained GAIN model, a personalized lifestyle recommendation approach is then proposed to identify the most suitable lifestyle modification for each risky individual, which

consists of two phases: (1) generate feasible alternative lifestyles (Sec. 4.2.2); and (2) select the optimal personalized lifestyle modification. (Sec. 4.3).

4.2.2 Phase 1: Generate feasible alternative lifestyles

As demonstrated in Figure 6, this phase consists of two steps, which aim to generate and filter the feasible alternative lifestyles. In lifestyle generation $\mathbf{X}_I$ is excluded since the lifestyle of the individual wanted to be generated based on characteristics. Thus, the first step tries to generate various potential alternative lifestyles, i.e., $\widehat{\mathbf{X}}_D$, using the trained GAIN model based on the given $\mathbf{X}_U$ of each individual. Subsequently, for the second step, the goal is to filter out the infeasible modifications, since to be counted as feasible, the alternative lifestyles have to satisfy the clinical guidelines (see details APPENDIX 1). For example, according to ref. [40, 41], the BMI value for a healthy lifestyle should not be more than 25 otherwise, the individual should try to reduce the weight, More specifically, increasing the BMI to 25 or higher cannot be treated as a feasible modification. Also, if the current BMI is higher than 25, then the suggested BMI cannot be higher than the current value in a feasible modification. Consequently, the generated alternative lifestyles that fit all the clinical guidelines will be selected as feasible lifestyles and the optimal one will be further identified in phase 2.

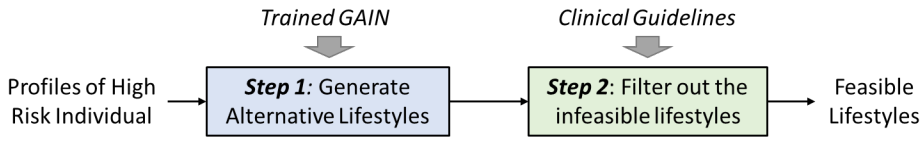

Figure 6: *The procedure to generate feasible alternative lifestyles (phase 1).*

4.3 Phase 2: Identify the optimal personalized lifestyle modification

Phase 2 aims to identify the optimal personalized lifestyle modification, which is the most suitable option for a risky individual by considering the CVD risk reduction, cost from the changes of lifestyle, and uncertainty in disease progression. As shown in Figure 7, this phase consists of three steps: 1) generate replicates (in terms of the indirectly changeable variables) for each feasible alternative lifestyles; 2) predict the CVD risk for each replicate; and 3) calculate the expected utility value for each feasible lifestyle and identify the optimal one.

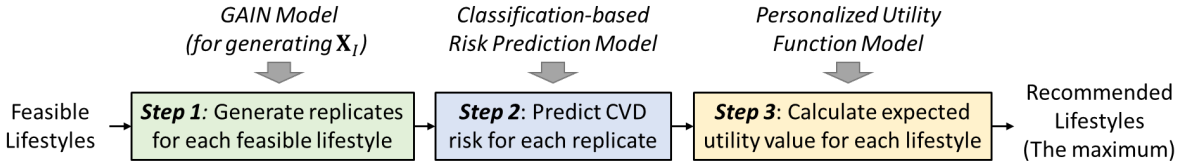

***Figure** 7: The procedure to identify the optimal personalized lifestyle modification in phase 2.*

Step 1 is focused on describing the uncertainty of disease progression under each feasible alternative lifestyle. In practice, the underlying relationship between $\mathbf{X}_I$ and $(\mathbf{X}_D, \mathbf{X}_U)$ contains significant uncertainties, which needs to be considered when evaluating the suggested alternative lifestyles. Therefore, for each feasible alternative lifestyle generated in phase 1, i.e., given $(\widehat{\mathbf{X}}_D, \mathbf{X}_U)$, generate N replicates for $\widehat{\mathbf{X}}_I$ using the GAIN model. Consequently, if M feasible alternative lifestyles are generated, each of them can be denoted as $\{{}_m^j\widehat{\mathbf{X}}\} = \{({}_m\widehat{\mathbf{X}}_D, {}_m^j\widehat{\mathbf{X}}_I, \mathbf{X}_U), j = 1,2,\cdots,N\})$, where $m = 1,2,\cdots,M$.

Subsequently, in step 2, the CVD risk for each ${}_m^j\widehat{\mathbf{X}}$ can be further predicted using the trained random forest model $\hat{f}$ (see more details in Section 4.1), i.e., ${}_m^j r = \hat{f}({}_m^j\widehat{\mathbf{X}})$. Then for a feasible alternative

lifestyle, the predicted overall CVD risk can be described by ${}_m\mathbf{r} = ({}_m^1r, {}_m^2r, \cdots, {}_m^Nr)$, where the uncertainty of disease progression is also quantified in risk estimation.

To further identify the optimal alternative lifestyle for different individuals, besides the overall level of risk, i.e., ${}_m\bar{r} = \frac{1}{N}\sum_{j=1}^{N} {}_m^jr$, the evaluation should also consider both the uncertainties of CVD risk and the cost of modifications (e.g., quit smoke, reduce BMI, etc.). Thus, it is also needed to develop an effective evaluation metric for decision-making. To achieve this goal, in step 3, a personalized assessment model, which is based on the expected utility theory, is proposed for optimal alternative lifestyle selection. In the proposed model, assessment is based on a developed personalized regularization-enabled exponential utility function, denoted as $u(r)$, which is demonstrated in Eq. (6),

$$u(r) = \frac{1 - e^{-\lambda(\mathbf{X}_D)[1-r(\mathbf{X}_D)]}}{\lambda(\mathbf{X}_D)} \tag{6}$$

where λ is the metric of uncertainty preference (larger λ shows less uncertainty averse) and r represents the estimated CVD risk for the lifestyle $\mathbf{X}_D$. In the traditional exponential utility function, λ is a pre-defined constant for each individual. However, in this study, different $\mathbf{X}_D$ will lead to different costs of lifestyle modification. Thus, even for the same individual, the value of λ may still vary since the cost of lifestyle modification could impact an individual's preference.

Consequently, the proposed model further extends the constant λ to a function of $\mathbf{X}_D$, denoted as $\lambda(\mathbf{X}_D)$, which considers the uncertainties from two aspects, namely, the disease progression (i.e, the variance of r) and the cost of lifestyle modifications. The key novel of $\lambda(\mathbf{X}_D)$ is introducing a personalized regularization term based on the cost of lifestyle modifications. Specifically, given an individual's original lifestyle $\mathbf{Z}_0 = [\mathbf{z}_0^1, \mathbf{z}_0^2, \cdots, \mathbf{z}_0^{11}]$, and the cost to change each directly changeable variable, a_i, $i = 1, \ldots, 11$, the total cost of lifestyle modification can be written as

$\sum_{i=1}^{11} |a_i(\mathbf{x}_D^i - \mathbf{z}_0^i)|$. Then by applying this cost to λ as a regularization term, $\lambda(\mathbf{X}_D)$ can be formulated as,

$$\lambda(\mathbf{X}_D) = \lambda_0 + \sum_{i=1}^{11} |a_i(\mathbf{x}_D^i - \mathbf{z}_0^i)| \tag{7}$$

where the first component, λ_0, represents the uncertainty preference of disease progression, which plays the same role as the λ in the traditional exponential utility function. The second component, i.e., the proposed regularization term, considers the cost of lifestyle modifications. A large modification cost will increase the value of λ and hence reduce the willingness to change.

Then based on the expected utility theory, the optimal $\mathbf{X}_D^*$ should maximize the expectation of $u(r)$, i.e., $\max_{\mathbf{X}_D} E[u(r)]$. In practice, $E[u(r)]$ can be estimated based on the averaged utility function value of each replicate, i.e., $\frac{1}{N}\sum_{j=1}^{N} u({}^j r|\mathbf{X}_D)$. Therefore, the selection of optimal $\mathbf{X}_D^*$ can be formulated by an optimization problem, as shown in Eq. (8),

$$\begin{aligned} &\max_{\mathbf{X}_D} \frac{1}{N} \sum\nolimits_{j=1}^{N} \frac{1 - e^{-\lambda(\mathbf{X}_D)\left[1 - {}^j r(\mathbf{X}_D)\right]}}{\lambda(\mathbf{X}_D)} \\ &s.t.\ \ \lambda(\mathbf{X}_D) = \lambda_0 + \sum_{i=1}^{11} |a_i(\mathbf{x}_D^i - \mathbf{z}_0^i)| \end{aligned} \tag{8}$$

where ${}^j r(\mathbf{X}_D)$ represents the predicted disease risk for the jth replicate under lifestyle $\mathbf{X}_D$. For a specific individual, from the M feasible alternative lifestyles (each has N replicates) generated from the GAIN model, the optimal $\mathbf{X}_D^*$ can be thereby selected through Eq. (8).

4.4 Performance validation

For the proposed method, it is also critical to validate its effectiveness. However, it is challenging since there is no actual patients are tested with a comparison target. Therefore, alternative, this

study aims to conduct the performance validation through three aspects: (1) the classification accuracy for risk prediction (Sec.4.4.1); (2) the effectiveness of the generated indirectly changeable variables (Sec.4.4.2); and (3) the effectiveness of the recommended alternative lifestyles for risk reduction (Sec.4.4.3).

4.4.1 Validation for the classification-based risk prediction

To validate the classification performance, the data are randomly separated into two parts, one training set (50%) and one testing set (50%). The classifiers are first trained using the training set, and then the testing set is used for validation. The classification performances are measured by the receiver operating characteristic (ROC) curve and Precision-Recall curve with its AUC scores.

4.4.2 Validation for the effectiveness of the generated indirectly changeable variables

Validating the effectiveness of generated indirectly changeable variables also plays a very important role in this study. As demonstrated in Figure 8, the proposed validation framework will first select samples from the testing set. Then it will use the trained GAIN model to generate the indirectly changeable variables $\{ {}^{j}\widehat{\boldsymbol{X}}_I, j = 1,2,\cdots,N\}$ for each sample with N replicates conditional on the actual lifestyle $\mathbf{X}_D$ and characteristics $\mathbf{X}_U$. Subsequently, the corresponding risk $\{ {}^{j}r,\ j = 1,2,\cdots,N\}$ can be thereby predicted. If $\{ {}^{j}r\}$ is consistent with the risk that is predicted using $\mathbf{X}_D$ and $\mathbf{X}_U$ with the actual $\mathbf{X}_I$, i.e., r_a, through comparison, then it can be concluded that the generated $\{ {}^{j}\widehat{\boldsymbol{X}}_I\}$ is effective. Specifically, the consistency comparison between $\{ {}^{j}r\}$ and r_a can be

conducted through two aspects: (1) to compare if r_a is inside the range of $\{{}^j r\}$; and (2) for all the samples, to compare if the average of $\{{}^j r\}$ is highly correlated with the corresponding r_a.

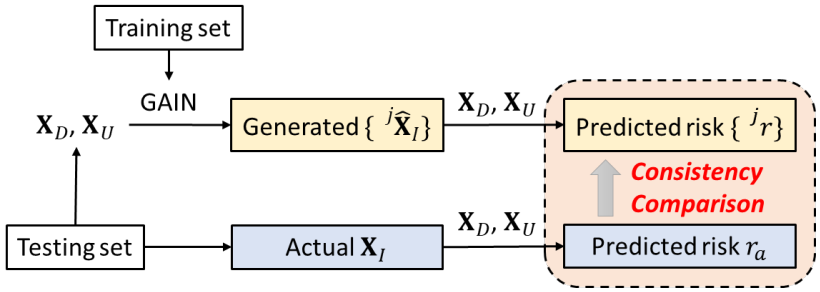

Figure 8: *The proposed framework to validate the effectiveness of the generated indirectly changeable variables.*

4.4.3 Validation for the effectiveness of the recommended lifestyle modification

Another critical validation in this study is the effectiveness of the recommended lifestyle modification for each individual. Essentially, the ultimate goal is to validate if the recommended lifestyle can significantly reduce CVD risk. Mathematically, it can be formulated as a hypothesis testing problem, i.e.,

$$\begin{aligned} H_0&: \bar{r}_{\text{rec}} = r_0 \\ H_1&: \bar{r}_{\text{rec}} < r_0 \end{aligned} \tag{9}$$

where r_0 represents the original CVD risk and $\bar{r}_{\text{rec}}$ represents the mean CVD risk after the recommended lifestyle modification. For a specific individual, r_0 can be assumed as a given constant (predicted by the trained RF model). Therefore, Eq. (9) can be implemented through t-test based on the generated $\{{}^j r_{\text{rec}}\}$, where $\{{}^j r_{\text{rec}}\}$ is the generated replicates under the recommended lifestyle modification.

Figure 9, the validation can be summarized by three steps: (1) pick up the high-risk individuals from the testing set; (2) predict the risk based on their current lifestyle and recommended lifestyle modification (with replicates), respectively; and (3) perform t test to see if $\bar{r}_{\mathrm{rec}}$ is significantly lower than r_0.

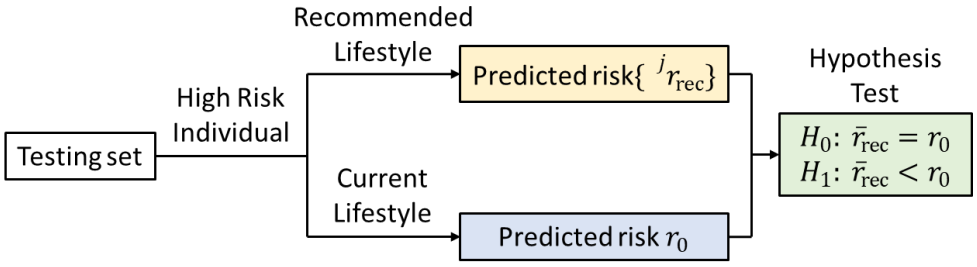

Figure 9: *The proposed framework to validate the effectiveness of the recommended lifestyle.*

CHAPTER V

RESULTS AND DISCUSSIONS

To demonstrate the effectiveness of the proposed methodology, the validation results are presented in this section. The results of classification performance are presented in Sec. 5.1. Sec. 5.2 validates the correctness of the indirectly changeable variables generated by GAIN. The effectiveness of the proposed personalized lifestyle recommendation approach is demonstrated in Sec. 5.3.

5.1 Classification results

To train and validate the classification models, 50% of the samples in the dataset are randomly selected for training, and the other 50% are used for the testing. Afterward, the comparison results, which are based on the ROC curve and AUC score, between three popular classification models, random forest, SVM, and k-NN classifiers, are presented in Figure 10. It can be observed that the random forest can provide superior classification performance than other methods for all outcome variables. Therefore, a random forest is selected in this study for CVD risk prediction.

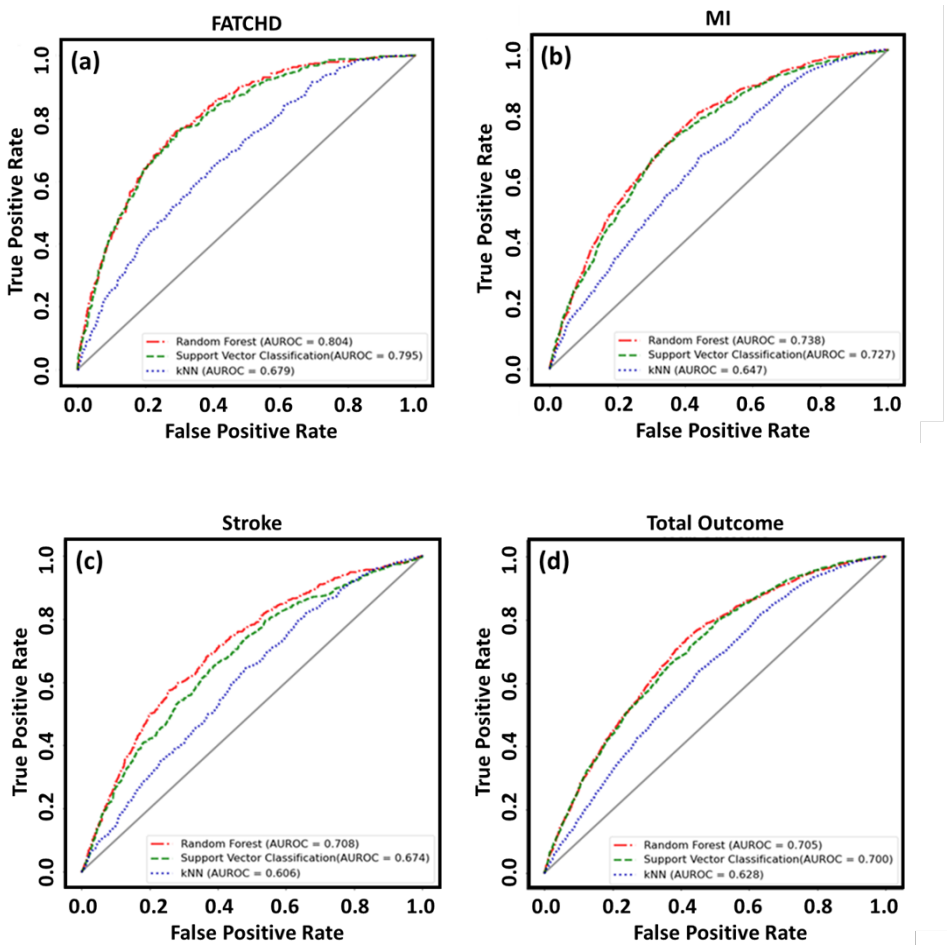

Figure 10: *ROC Curves based on different classification models for the four outcome variables. (a) Fatal CHD; (b) MI; (c) Stroke; and (d) Total Outcome that sum the other outcomes.*

Furthermore, the detailed classification results using random forest are presented in Table 4 (for Precision-Recall graph please see APPENDIX 2). For all of the four outcome variables, the AUC score can achieve at least 70%, which indicate that the trained classification model $\hat{f}$ is able to effectively predict the CVD risk of an individual based on its profile $\boldsymbol{X}$ (Eq. (4)). Besides that, 10-fold cross-validation is also applied, and the AUC scores are also very close.

Table 4: Data distributions and AUC Scores of the Outcomes

CVD Type	Sample Size	Diagnosed Patient Size	Diagnosed Patient Rate	AUC Score	AUC Score (Cross Validation)
Fatal CHD	10744	677	0.063	0.82	0.81
MI	11635	1538	0.132	0.74	0.72
Stroke	11292	1195	0.106	0.70	0.70
Total Outcome	13654	3557	0.260	0.70	0.69

5.2 Evaluation of the generated indirectly changeable variables

This section presents the evaluation results for the generated indirectly changeable variables by using the GAIN algorithm. The parameters are identified as shown in Table 5 for this experiment.

Table 5: The values for certain parameters of GAIN

Parameters	Value
batch_size	32
hint_rate	0.5
alpha	1
iterations	5000

To conduct the validation with this algorithm, 200 of the truly healthy individuals who are correctly classified as healthy, and 200 of the truly unhealthy individuals who are correctly classified as unhealthy are randomly selected from the testing set (50% samples, see Sec. 4.1) are selected as input. As presented in Sec. 3.4.2, for each individual, the indirectly changeable variables with N replicates, i.e., $\{ {}^{j}\widehat{\boldsymbol{X}}_I, j = 1,2,\cdots,N\}$, were generated conditional on the individual's actual lifestyle $\boldsymbol{X}_D$ and characteristics $\boldsymbol{X}_U$. Then the corresponding risk $\{ {}^{j}r,\ j = 1,2,\cdots,N\}$ for each replicate was predicted as well. Notably, the number of replicates (N) is set to 500. As presented in Figure 11, the Pearson correlation coefficient between the averaged $\{ {}^{j}r\}$ and the corresponding r_a can

achieve 91% in Fatal CHD dataset. In the other datasets; MI, Stroke, and Total Outcome this value follows with 82%, 95%, and 80% respectively as shown. Overall it can be said that the Pearson correlation coefficient is greater than 80%, which indicates that the generated indirectly changeable variables using GAIN can well approximate to actual situations of the datasets.

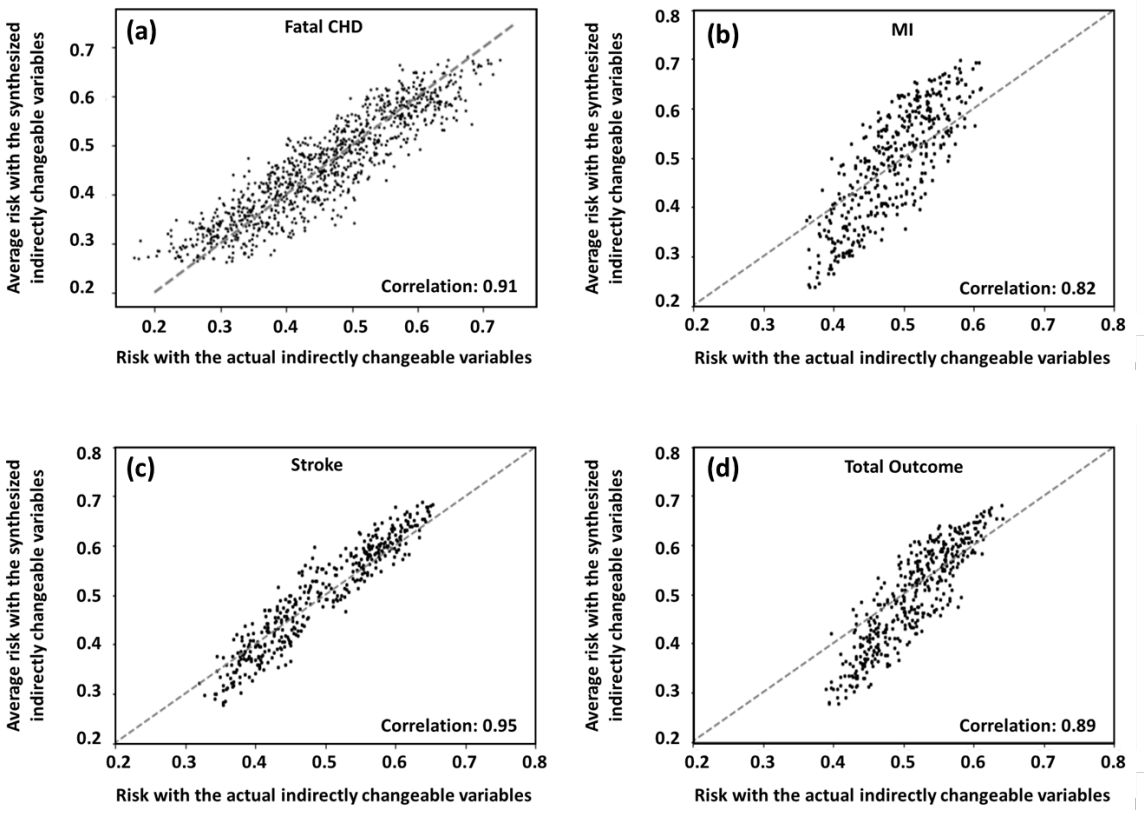

Figure 11: *The correlation between average risk with the synthesized indirectly changeable variables vs. risk with actual indirectly changeable variables. (a) Fatal CHD; (b) MI; (c) Stroke; (d) Total outcome.*

Moreover, 10 individuals were randomly selected from the data used for Figure 12, and then further compare if r_a is inside the range of $\{{}^{j}r\}$ for each individual. As demonstrated in Figure 12, the light blue boxes represent the range of $\{{}^{j}r\}$, and the dark blue line inside each box represents the

value of r_a. From the results it can be observed the range of the $\{ {}^{j}r\}$ can well cover the value of r_a, which further indicates that the GAIN-based generation for $\mathbf{X}_I$ is effective.

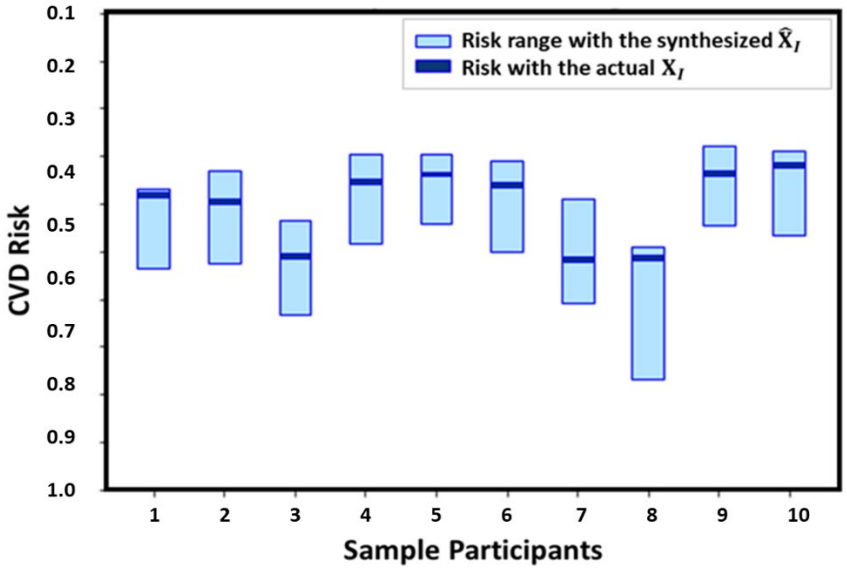

Figure 12: *Risk range of the replicates after synthesizing the indirectly changeable features by GAIN and the original risk with original indirectly changeable features for 10 randomly selected participants.*

5.3 Evaluation of the suggested lifestyle modifications

As introduced in Sec. 3.4.3, this section demonstrates the validation results for the recommended lifestyle modifications using a hypothesis test. Again the parameters that are shown in Table 5 are used for the GAIN imputation algorithm. To avoid the potential bias invalidation, the truly unhealthy individuals who are correctly classified as unhealthy from the testing set (50% samples, see Sec. 4.1) which indicates 285, 644, 513, 1492 participants from fatal CHD, MI, Stroke and General Outcome datasets are selected as testing datasets, respectively. To identify the optimal lifestyle modification for all of these individuals, the number of replicates (N) for each feasible modification is set to 500. Other than that, for the utility function that is applied to the optimal

lifestyle selection, the parameters of $\lambda(\mathbf{X}_{\mathrm{D}})$ are defined as same for all the individuals in this experiment. First, the λ_0 is defined as 0 which is the lowest risk aversion. Then, the $a_i,\ i = 1, \ldots, 11$ values shown in Table 6.

Table 6: The coefficients of indirectly changeable features

Variables	i	a_i
BMI01	1	0.01
ETHANL03	2	0.0005
SMKSTA	3	0.001
EXE01	4	0.001
CARB	5	0.0005
CHOL	6	0.001
DFIB	7	0.001
PROT	8	0.0005
SFAT	9	0.0005
TFAT	10	0.0005
TCAL	11	0.0005

For instance, a unit change on the discretized category of the BMI01 variable in the alternative lifestyle will be multiplied with $a_i = 0.01$. For this experiment changing the weight is defined as the lifestyle change which requires the most effort, so the highest coefficient is used for BMI.

For fatal CHD (y_1), the hypothesis test results in terms of p-value are presented in Figure 13, which shows that the p-values for 98% individuals are lower than 0.05, and only 1.6% are larger than 0.1. Therefore, the validation results demonstrate the proposed method can significantly reduce the potential fatal CHD risk of risky individuals.

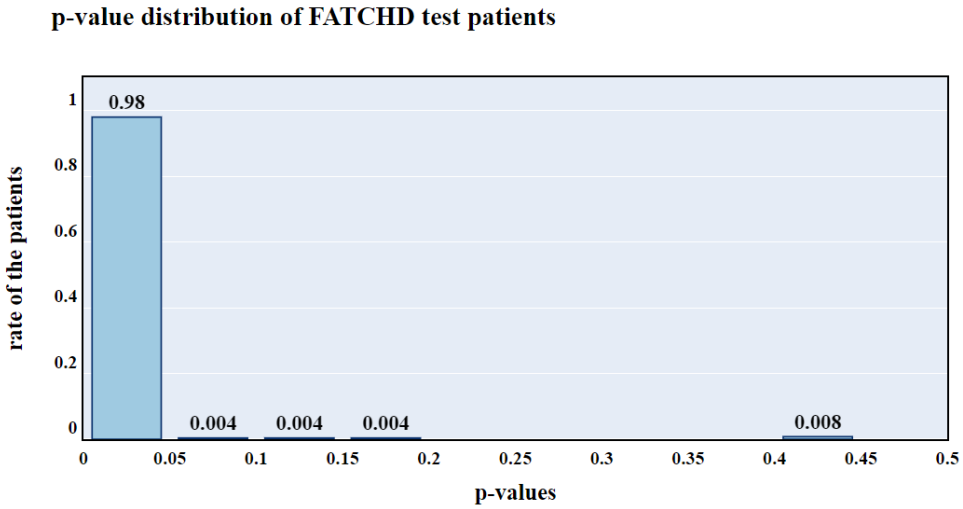

Figure 13: *Hypothesis testing of risk reduction with recommended lifestyles for fatal CHD:*

$$H_0: \bar{r}_{\text{rec}} = r_0, H_1: \bar{r}_{\text{rec}} < r_0$$

Table 7 summarizes the comparison between the risk of the original and the recommended lifestyles for the patients in the fatal CHD dataset. In this table, the "Original" column displays the average risk of original lifestyle, "Recommended" column displays the average risk of recommended lifestyle diagnosing with a fatal CHD in 10 years for test patients. Each individual got only 1 type of recommended lifestyle based on their preferences. The row "Avg relative risk reduction" displays the average reduction with the selected recommended lifestyle for each of the test patients with the standard deviation in parenthesis. Besides that, the last row "Avg p-value" displays the average p-value of the hypothesis testing result; $H_0: \bar{r}_{\text{rec}} = r_0$, $H_1: \bar{r}_{\text{rec}} < r_0$ for each testing patient. Therefore, using large data (more than 10000 participants) and a significant reduction in the risk with recommended lifestyles caused the small p-value. The results also demonstrate the proposed method can significantly reduce the potential CVD risk for the test patients. In detail, it can be said that the average risk of fatal CHD can be reduced by 9.4%.

Table 7: Comparison of original and the predicted fatal CHD test patients

Fatal CHD	Original	Recommended
Avg risk	55.6%	46.1%
Avg relative risk reduction	-	9.4% (5.19%)
Avg p-value	-	0.00491

For MI (y_2), the hypothesis test results in terms of p-value are presented in Figure 14, which shows that the p-values for 98.3% individuals are lower than 0.05, and only 1% are larger than 0.1. Therefore, the validation results demonstrate the proposed method can significantly reduce the potential MI risk of risky individuals.

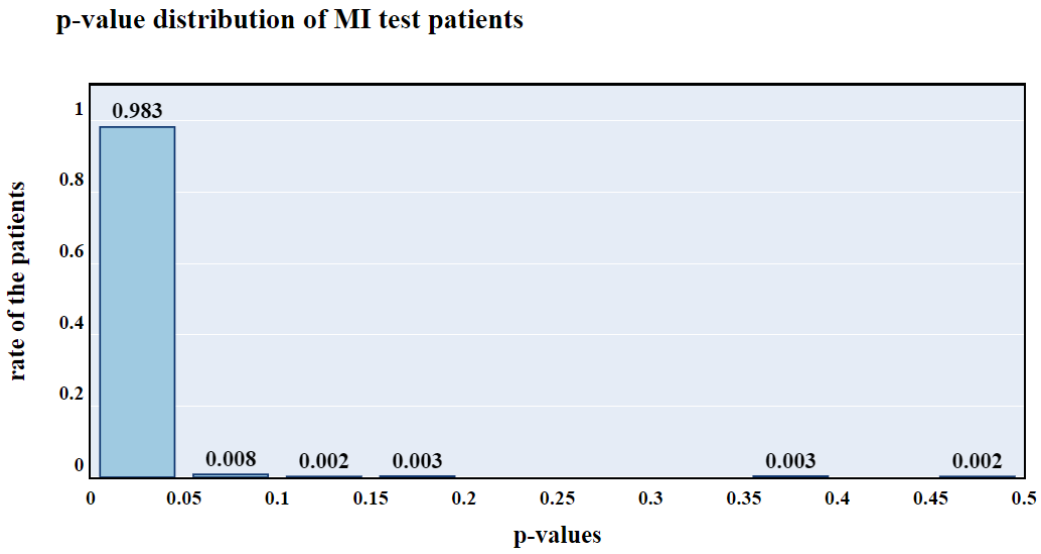

Figure 14: *Hypothesis testing of risk reduction with recommended lifestyles for MI test patients:*

$$H_0: \bar{r}_{\text{rec}} = r_0, H_1: \bar{r}_{\text{rec}} < r_0$$

Table 8 summarizes the comparison of the risks of diagnosing with MI in 10 years with original and the recommended lifestyles for the patients. In this table, the "Original" column displays the average risk of original lifestyle, "Recommended" column displays the average risk of

recommended lifestyle diagnosing with a MI in 10 years for test patients. Each individual got only 1 type of recommended lifestyle based on their preferences. Based on the table, it can be said that again using large data and a significant reduction in the risk with recommended lifestyles caused the small p-value. Thus, the risk of diagnosing with MI in 10 years can be lowered by 8.65%.

Table 8: Comparison of original and the predicted MI test patients

MI	Original	Recommended
Avg risk	56.0%	47.3%
Avg relative risk reduction	-	8.65% (4.58%)
Avg p-value	-	0.00326

For Stroke (y_3), the hypothesis test results in terms of p-value are presented in Figure 15, which shows that the p-values for 99.2% individuals are lower than 0.05, and only 0.006% are larger than 0.1. Therefore, the validation results demonstrate the proposed method can significantly reduce the potential Stroke risk of risky individuals.

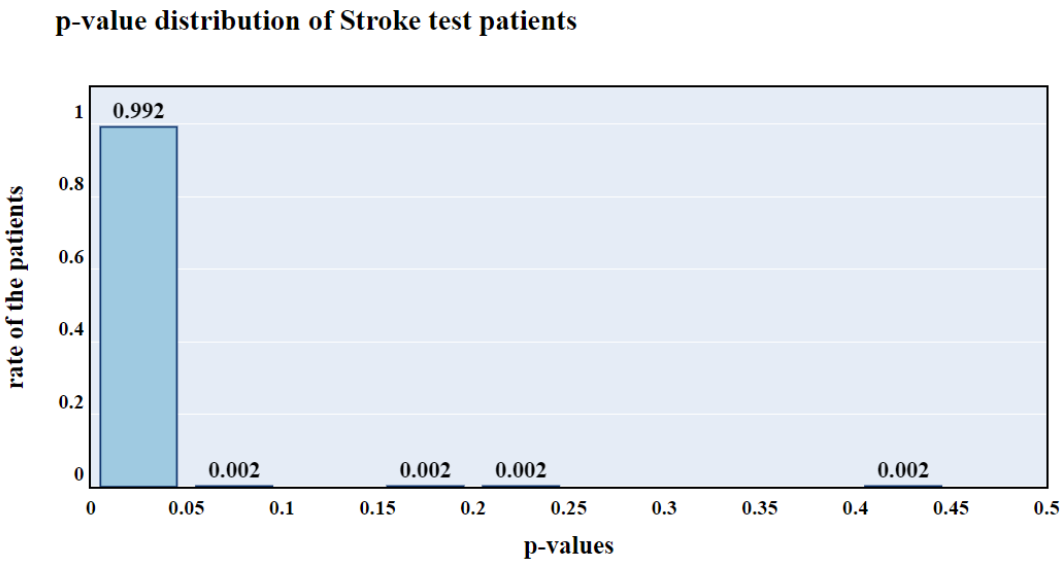

Figure 15: *Hypothesis testing of risk reduction with recommended lifestyles for Stroke test patients:* $H_0: \bar{r}_{rec} = r_0$, $H_1: \bar{r}_{rec} < r_0$

Table 9 summarizes the comparison of the risks of diagnosing with Stroke in 10 years with original and the recommended lifestyles for the patients. In this table, the "Original" column displays the average risk of original lifestyle, "Recommended" column displays the average risk of recommended lifestyle diagnosing with a Stoke in 10 years for test patients. Each individual got only 1 type of recommended lifestyle based on their preferences. Based on the table, it can be said that again using large data (more than 10000 participants) and a significant reduction in the risk with recommended lifestyles caused the small p-value. Therefore, the risk of diagnosing with Stoke in 10 years can be lowered by 5.3%.

Table 9: Comparison of original and the predicted Stroke test patients

Stroke	Original	Recommended
Avg risk	54.2%	48.8%
Avg relative risk reduction	-	5.3% (2.93%)
Avg p-value	-	0.00499

For general outcome (y_3), which is the combination of fatal CHD, MI, and Stroke, the hypothesis test results in terms of p-value are presented in Figure 16, which shows that the p-values for 98.8% individuals are lower than 0.05, and only 0.01% are larger than 0.1. Therefore, the validation results demonstrate the proposed method can significantly reduce the potential generally CVD risk of risky individuals.

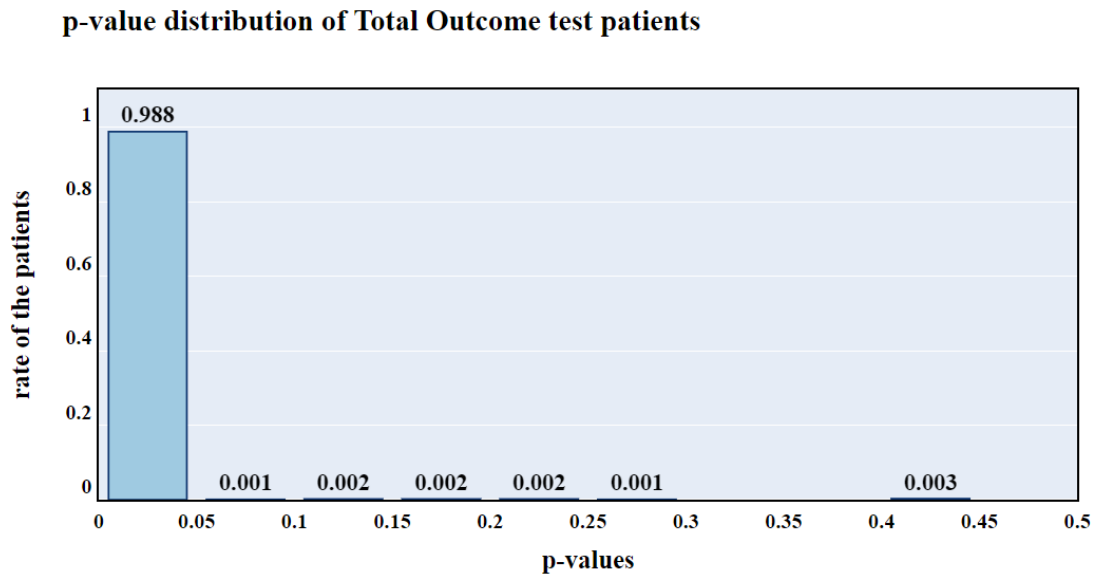

Figure 16: *Hypothesis testing of risk reduction with recommended lifestyles for Total Outcome test patients:* $H_0: \bar{r}_{rec} = r_0$, $H_1: \bar{r}_{rec} < r_0$

Table 10 shows the comparison between the risk of the original and the recommended lifestyles for total outcomes. In this table, the "Original" column displays the average risk of original lifestyle, "Recommended" column displays the average risk of recommended lifestyle diagnosing with one of the CVD types from fatal CHD, MI, or Stroke in 10 years for test patients. Each individual got only 1 type of recommended lifestyle based on their preferences. Based on the table, it can be said that again using large data (more than 10000 participants) and a significant reduction in the risk with recommended lifestyles caused the small p-value. The risk of diagnosing with fatal CHD, MI,

or Stroke in 10 years can be lowered by 5.7% which is a lower reduction compared to the previous dataset applications. The rate of unhealthy patients is higher in the dataset compare to the previous datasets since it is a combination of all.

Table 10: Comparison of original and the predicted Total Outcome test patients

Total Outcome	Original	Recommended
Avg risk	55.6%	49.8%
Avg relative risk reduction	-	5.7% (3.29%)
Avg p-value	-	0.00285

Based on these tables it can be interpreted that the risk reductions are higher when the target is more clear for the algorithm. However, aiming multiple CVD types together to reduce the reduction might return a lower average relative risk reduction. For instance, when patients are provided the lifestyle recommendations for just Fatal CHD and just MI the average risk reduction is 9.4% and 8.65% respectively. When the patients are asked to lifestyle improvements for Stroke and Total outcome the average risk reduction becomes 5.3% and 5.7% respectively. Since it is mentioned above that Stroke is a combination of the ischemic, incident, or hemorrhagic types of Stoke, and total outcome is the combination of all the types fatal CHD, MI, and Stroke. Based on the results it can be seen that investigating CVD types separately can cause better risk reductions for that type. However, when more than one type is included risk, the risk can be reduced against all the types as lower. Here, based on the expert and the patient preferences choices can be made while making a recommendation to the individual.

Other than this general expression, Figure 17 demonstrated the risk reduction with the recommended lifestyles for a randomly selected high Fatal CVD risk individual, Patient 49. It illustrates the recommendations for the patient to decrease the risk of diagnosing with fatal CHD

in 10 years. For instance, it suggests a reduction in BMI, saturated fatty acid (g/day), and cholesterol (mg), while keeping the alcohol (g/day), dietary fiber, total fat (g/day), and smoking status the same. Besides that, the proposed method recommends to this patient to increase the physical activity (h/week), protein intake (g), total energy (mg/dL), and carbohydrate (g). These results can be interpreted as that the two-phase framework suggests to the patient to change the resource of the energy. Replace the energy that is coming from saturated fatty acid and cholesterol with the energy from other carbohydrates, protein, and increase the time for exercising. The random forest classifier predicts that with the application of these recommendations the patient can reduce the original risk of diagnosing with Fatal CHD in 10 years by 19%.

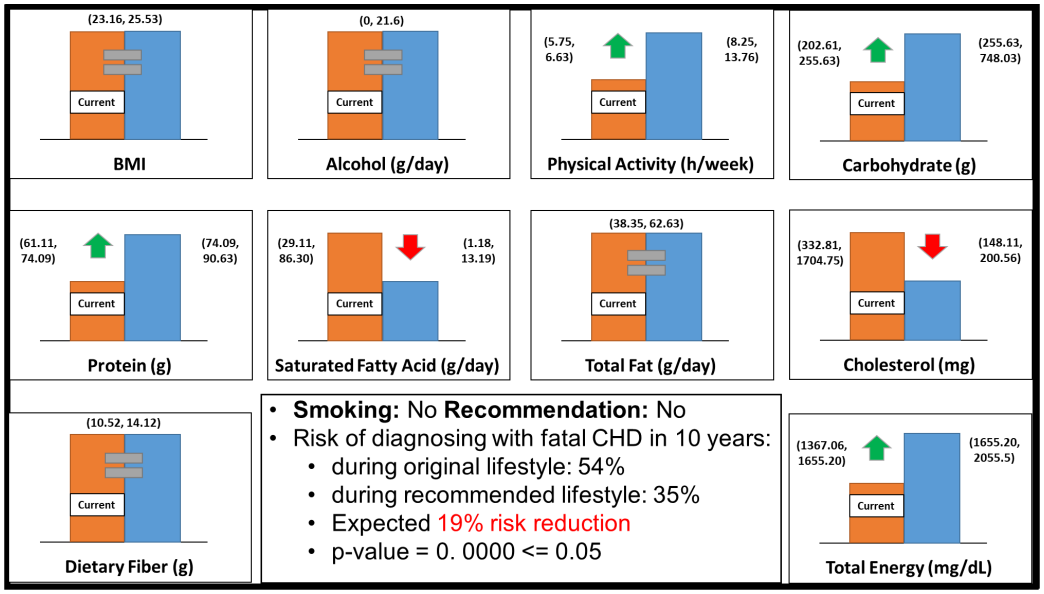

***Figure 17**: Lifestyle recommendations for one of the randomly selected fatal CHD patients.*

Similarly, Figure 18 illustrates the recommendations for Patient 156 to decrease the risk of diagnosing with Stroke in 10 years. This time, it suggests a reduction in BMI, carbohydrate(g), cholesterol (mg), and total energy (mg/dL) while keeping the alcohol (g/day), dietary fiber (g),

physical activity (h/week), and smoking status the same. Besides that, the proposed method recommends to this patient to increase the protein intake (g), total fat (mg/dL), and saturated fatty acid (g/day). From this figure, it can be interpreted as the two-phase framework recommending to the patient to reduce total energy intake and increase the protein consumption and fat resources at a healthy level to reduce the risk. Even though the framework suggests increasing fat consumption, this increase does not go above the literature guidelines (see APPENDIX 1). The random forest classifier predicts that with the application of these recommendations the patient can reduce the original risk of diagnosing with Stroke in 10 years almost by 5%.

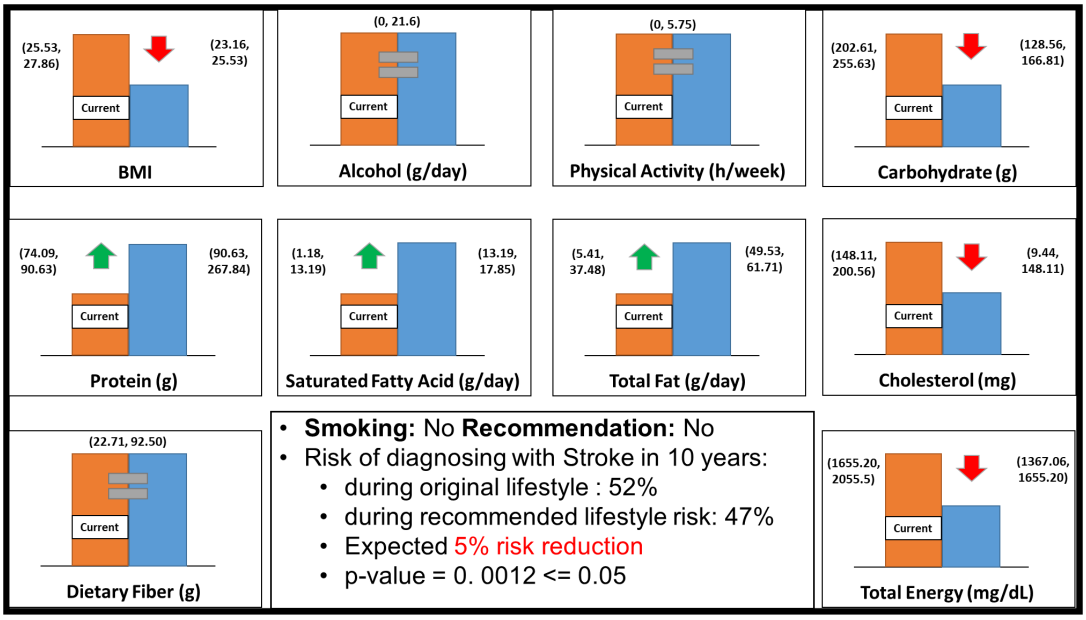

Figure 18: *Lifestyle recommendations for one of the randomly selected Stroke patients.*

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

This research established a data-driven personalized lifestyle recommendation framework by incorporating random forest classification, the GAN model, and the expected utility hypothesis. The recommended lifestyle modification is the potential to effectively reduce the risk of common CVDs for each individual. The proposed methodology can be applied as a decision-support tool for healthcare professionals and high-risk individuals in conducting an optimal applicable lifestyle by learning from the related healthcare data. The aim of this proposed model is the obtaining the best match between recommended lifestyles, patient's preferences, and limitations. The predictive model captures the relationships between the recommendations, constraints, and CVD risk outcomes. The selection of the optimum lifestyles is made by these parameters like it is also explained in Section 4.3.

In practice, to make the recommendation scenario more realistic, we can also further consider more CVD-related risk factors. For example, the socio-economic factors, including fast food options, availability of fresh and healthy food, time availability to increase exercise timing, etc. With more related variables, it is possible to further increase the accuracy of our proposed method. Besides, genetic factors such as family genetic history might be also helpful to improve the recommendation performance.

Meanwhile, it may be also helpful to jointly consider the data from multiple visits. For example, the year-by-year longitudinal patterns, e.g., change of diet habits, medication history, could be learned by the machine learning algorithms and utilized for offering better recommendation options. Therefore, more machine learning algorithms that are capable of handling longitudinal patterns could be adopted in future work.

The different philosophies exist in lifestyle studies with statistical methods. Such as the ARIC and Framingham models have some different elements compare to the machine learning methods since the purpose of them mostly on variable relationships and causal effects on output. The significant risk factors to explain most of the CVD risks in the population the variables that have been detected and used in ARIC and Framingham are fair enough. Compared with the existing lifestyle studies in CVD, this study includes more variables, so that more individualized information can be utilized for risk assessment and lifestyle recommendation. Furthermore, incorporating the powerful machine learning algorithms and utility model also enables this study to significantly improve the predictive modeling accuracy as well as the recommendation effectiveness and flexibility. In the proposed method, the GAIN algorithm learns the patterns of healthy participants' lifestyles and the complex interactions between variables. Then some feasible alternative lifestyles can be generated, and also the expected indirectly changeable variables for these generated alternative lifestyles can be predicted by GAIN as well. With the help of the proposed utility function, the optimal can be selected by leveraging the CVD risk, the patient's preferences, and practical restrictions. machine learning focus on the prediction more compares to the statistical method. Using more variables that provide more information to increase predictive performance is a common way. Thus, this study includes more variables compares to the ARIC and Framingham studies. Since the difference between interpretation and prediction adding more variables is not causing big problems. Additionally, the GAIN algorithm can learn the interactions of the variables with its highly

randomized structure. Thus, it can provide many different combinations of lifestyle recommendations that are parallel to the literature guidelines.

On the other hand, the statistical method that is applied in ARIC and Framingham studies are optimized their results for the majority of the sample size. However, the recommendations based on the lifestyles may vary for different people. The lifestyle recommendation for a 180 lb dancer who exercises hours in a day can not be the same with a 180 lb academician who sits in office hours in a day. Because of these obvious differences in population, the proposed algorithm of this study provides personalized lifestyle recommendations by optimizing them based on the predefined constraints. The participation of the patients in the treatment process can be increased with this interactive solution method. This way can also cause improvement in the perseverance of the patients to get healthy lifestyles since they can face the cause and effect of their lifestyle and attitude directly.

In this proposed method, the GAIN algorithm learns from the healthy participants' lifestyles and complex variable interactions between lifestyles and characteristics of them. Then for the test patients, the lifestyles are generated with many healthy possibilities with the help of the clinical guidelines filter. The indirectly changeable features are also predicted by GAIN based on the healthy participants' variable interactions. From these many possibilities, the optimum one is selected based on CVD risk, the patient's preferences, and limitations.

In summary, this study developed a data-driven personalized lifestyle recommendation framework for CVD prevention. A proof-of-concept case study based on the ARIC dataset demonstrated that our method is able to provide appropriate individualized lifestyle recommendations and thereby effectively reduce the CVD risk. The proposed method is promising to enable better use of individual-level healthcare data for more effective CVD prevention. This study is also very promising to be extended to more types of diseases (e.g., hypertension, diabetes, etc.) and

contribute to the researches related to behavior change, personalized treatment, and medicine. In future work, more related data and emergy machine learning techniques will be effectively incorporated to further improve the recommendation accuracy and capability.

REFERENCES

1. Benjamin, E.J., et al., *Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association.* Circulation, 2019. **139**(10): p. e56-e528.
2. Wawrzyniak, A.J., *Framingham Heart Study*, in *Encyclopedia of Behavioral Medicine*, M.D. Gellman and J.R. Turner, Editors. 2013, Springer New York: New York, NY. p. 811-814.
3. McGill, H.C., Jr., C.A. McMahan, and S.S. Gidding, *Preventing heart disease in the 21st century: implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) study.* Circulation, 2008. **117**(9): p. 1216-27.
4. Atherosclerosis Risk in Communities (ARIC) Study, *The project description and data.*
5. Goodfellow, I., et al. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.
6. Eaker, E.D. and W.P. Castelli, *Coronary heart disease and its risk factors among women in the Framingham Study*, in *Coronary heart disease in women*. 1987, Haymarket Doyma, New York. p. 122-130.
7. Vasan, R.S., et al., *Antecedent blood pressure and risk of cardiovascular disease: the Framingham Heart Study.* Circulation, 2002. **105**(1): p. 48-53.
8. Wilson, P.W., W.P. Castelli, and W.B. Kannel, *Coronary risk prediction in adults (the Framingham Heart Study).* The American journal of cardiology, 1987. **59**(14): p. G91-G94.
9. Millen, B.E., et al., *Dietary patterns and the odds of carotid atherosclerosis in women: the Framingham Nutrition Studies.* Preventive medicine, 2002. **35**(6): p. 540-547.
10. Pencina, M.J., et al., *Predicting the 30-Year Risk of Cardiovascular Disease.* Circulation, 2009. **119**(24): p. 3078-3084.
11. LeFevre, M.L., *Behavioral counseling to promote a healthful diet and physical activity for cardiovascular disease prevention in adults with cardiovascular risk factors: US Preventive Services Task Force Recommendation Statement.* Annals of internal medicine, 2014. **161**(8): p. 587-593.
12. Buttar, H.S., T. Li, and N. Ravi, *Prevention of cardiovascular diseases: Role of exercise, dietary interventions, obesity and smoking cessation.* Experimental and clinical cardiology, 2005. **10**(4): p. 229-249.
13. Chi, C.-L., et al., *Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options.* Journal of Biomedical Informatics, 2012. **45**(6): p. 1164-1174.
14. Mansoor, H., et al., *Novel Self-Report Tool for Cardiovascular Risk Assessment.* Journal of the American Heart Association, 2019. **8**(24): p. e014123.
15. Chambless, L.E., et al., *Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study.* Journal of Clinical Epidemiology, 2003. **56**(9): p. 880-890.
16. Mansoor, H., et al., *Novel Self-Report Tool for Cardiovascular Risk Assessment.* J Am Heart Assoc, 2019. **8**(24): p. e014123.
17. Enwald, H.P.K. and M.-L.A. Huotari, *Preventing the obesity epidemic by second generation tailored health communication: an interdisciplinary review.* Journal of medical Internet research, 2010. **12**(2): p. e24.

18. Skinner, C.S., et al., *How effective is tailored print communication?* Annals of Behavioral Medicine, 1999. **21**(4): p. 290-298.
19. Kreuter, M.W. and R.J. Wray, *Tailored and targeted health communication: strategies for enhancing information relevance.* American journal of health behavior, 2003. **27**(1): p. S227-S232.
20. Brug, J., A. Oenema, and M. Campbell, *Past, present, and future of computer-tailored nutrition education.* The American Journal of Clinical Nutrition, 2003. **77**(4): p. 1028S-1034S.
21. Oenema, A., F. Tan, and J. Brug, *Short-term efficacy of a web-based computer-tailored nutrition intervention: main effects and mediators.* Annals of behavioral medicine, 2005. **29**(1): p. 54-63.
22. Jabeen, F., et al., *An IoT based efficient hybrid recommender system for cardiovascular disease.* Peer-to-Peer Networking and Applications, 2019. **12**(5): p. 1263-1276.
23. Nam, Y. and Y. Kim, *Individualized exercise and diet recommendations: an expert system for monitoring physical activity and lifestyle interventions in obesity.* Journal of Electrical Engineering & Technology, 2015. **10**(6): p. 2434-2441.
24. Kang, W.-C., et al. *Visually-aware fashion recommendation and design with generative image models*. in *2017 IEEE International Conference on Data Mining (ICDM)*. 2017. IEEE.
25. He, X., et al. *Adversarial personalized ranking for recommendation*. in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018.
26. Li, G., et al., *Asking images: Hybrid recommendation system for tourist spots by hierarchical sampling statistics and multimodal visual Bayesian personalized ranking.* IEEE Access, 2019. **7**: p. 126539-126560.
27. Chen, X., et al., *Generative Adversarial User Model for Reinforcement Learning Based Recommendation System*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 1052--1061.
28. Bharadhwaj, H., H. Park, and B.Y. Lim, *RecGAN: recurrent generative adversarial networks for recommendation systems*, in *Proceedings of the 12th ACM Conference on Recommender Systems*. 2018, Association for Computing Machinery: Vancouver, British Columbia, Canada. p. 372–376.
29. Gao, R., et al. *DRCGR: Deep Reinforcement Learning Framework Incorporating CNN and GAN-Based for Interactive Recommendation*. in *2019 IEEE International Conference on Data Mining (ICDM)*. 2019. IEEE.
30. Chambless, L.E., et al., *Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the Atherosclerosis Risk in Communities (ARIC) Study, 1987-1993.* Am J Epidemiol, 1997. **146**(6): p. 483-94.
31. Lichtenstein, A.H., et al., *Diet and lifestyle recommendations revision 2006: a scientific statement from the American Heart Association Nutrition Committee.* Circulation, 2006. **114**(1): p. 82-96.
32. Whelton, S.P., et al., *Effect of aerobic exercise on blood pressure: a meta-analysis of randomized, controlled trials.* Annals of internal medicine, 2002. **136**(7): p. 493-503.
33. Stekhoven, D.J. and P. Bühlmann, *MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics, 2011. **28**(1): p. 112-118.
34. Beretta, L. and A. Santaniello, *Nearest neighbor imputation algorithms: a critical evaluation.* BMC Med Inform Decis Mak, 2016. **16 Suppl 3**: p. 74.
35. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
36. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.
37. Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.
38. Fawcett, T., *An introduction to ROC analysis.* Pattern recognition letters, 2006. **27**(8): p. 861-874.

39. Yoon, J., J. Jordon, and M.v.d. Schaar. *GAIN: Missing Data Imputation using Generative Adversarial Nets*. in *ICML*. 2018.
40. Nuttall, F.Q., *Body Mass Index: Obesity, BMI, and Health: A Critical Review.* Nutrition Today, 2015. **50**(3): p. 117-128.
41. Kuczmarski, R.J. and K.M. Flegal, *Criteria for definition of overweight in transition: background and recommendations for the United States.* The American Journal of Clinical Nutrition, 2000. **72**(5): p. 1074-1081.

APPENDICES

APPENDIX 1

The filters that are applied in the lifestyle recommendation based on clinical guidelines [1, 16, 31]:

BMI:

$$x_{\mathrm{BMI}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{BMI}}^{\mathrm{rec}} \leq 25 & \text{if } x_{\mathrm{BMI}}^{\mathrm{ori}} \leq 25 \\ x_{\mathrm{BMI}}^{\mathrm{rec}} \leq x_{\mathrm{BMI}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Alcohol intake:

$$x_{\mathrm{ETHANOL03}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{ETHANOL03}}^{\mathrm{rec}} \leq 196 & \text{if } x_{\mathrm{ETHANOL03}}^{\mathrm{ori}} \leq 196 \\ x_{\mathrm{ETHANOL03}}^{\mathrm{rec}} \leq x_{\mathrm{ETHANOL03}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Smoking:

$$x_{\mathrm{SMKSTA}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{SMKSTA}}^{\mathrm{rec}} = 0 & \text{if } x_{\mathrm{SMKSTA}}^{\mathrm{ori}} = 0 \\ x_{\mathrm{SMKSTA}}^{\mathrm{rec}} \leq x_{\mathrm{SMKSTA}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Cholesterol:

$$x_{\mathrm{CHOL}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{CHOL}}^{\mathrm{rec}} \leq 300 & \text{if } x_{\mathrm{CHOL}}^{\mathrm{ori}} \leq 300 \\ x_{\mathrm{CHOL}}^{\mathrm{rec}} \leq x_{\mathrm{CHOL}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Dietary fiber:

$$x_{\mathrm{DFIB}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{DFIB}}^{\mathrm{rec}} \geq 25 & \text{if } x_{\mathrm{DFIB}}^{\mathrm{ori}} \geq 25 \\ x_{\mathrm{DFIB}}^{\mathrm{rec}} \geq x_{\mathrm{DFIB}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Saturated fatty acids:

$$x_{\mathrm{SFAT}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{SFAT}}^{\mathrm{rec}} \leq 22 & \text{if } x_{\mathrm{SFAT}}^{\mathrm{ori}} \leq 22 \\ x_{\mathrm{SFAT}}^{\mathrm{rec}} \leq x_{\mathrm{SFAT}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

Total fats:

$$x_{\mathrm{TFAT}}^{\mathrm{rec}} = \begin{cases} x_{\mathrm{TFAT}}^{\mathrm{rec}} \leq 78 & \text{if } x_{\mathrm{TFAT}}^{\mathrm{ori}} \leq 78 \\ x_{\mathrm{TFAT}}^{\mathrm{rec}} \leq x_{\mathrm{TFAT}}^{\mathrm{ori}} & \text{otherwise} \end{cases}$$

APPENDIX 2

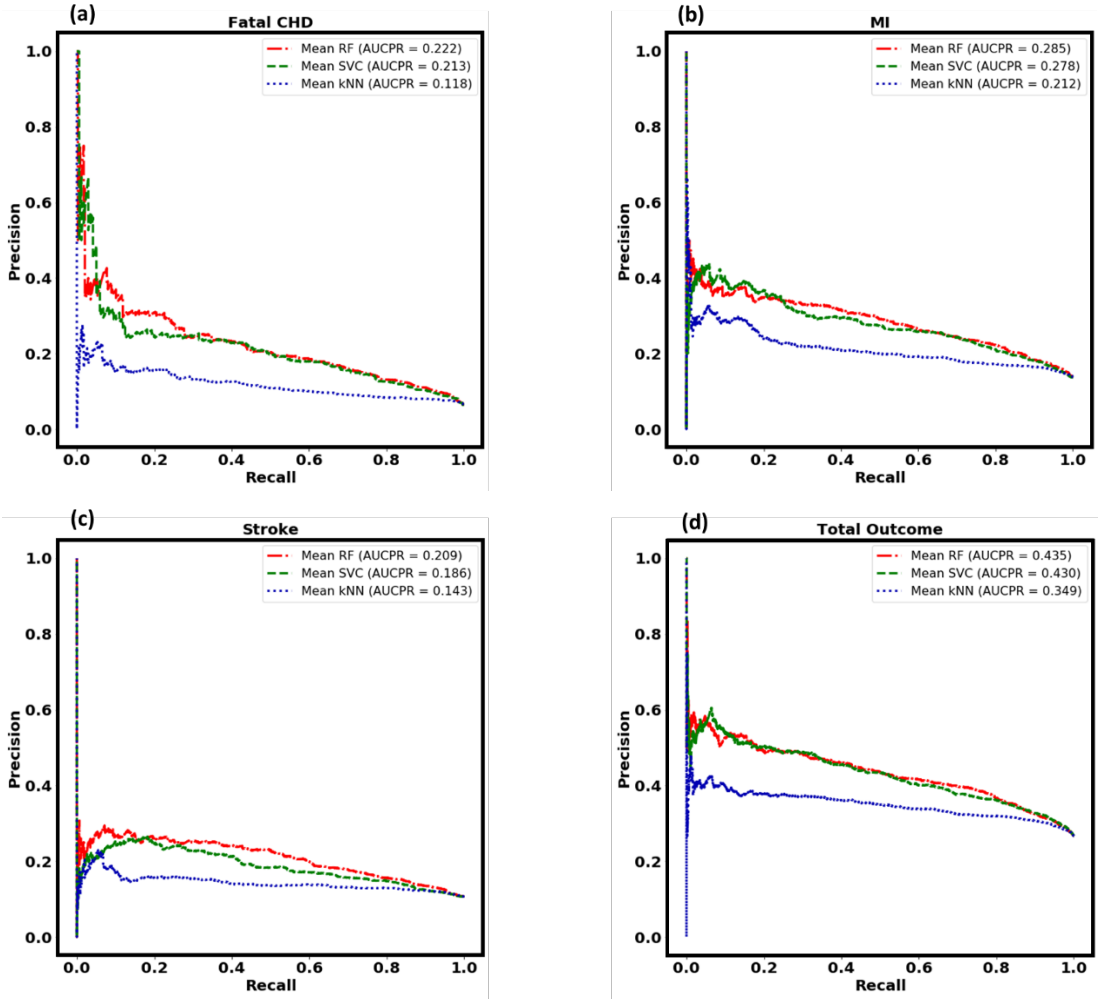

Figure 19: *Precision-Recall Curves based on different classification models for the four outcome variables. (a) Fatal CHD; (b) MI; (c) Stroke; and (d) Total Outcome that sum the other outcomes.*

VITA

Ayse Dogan

Candidate for the Degree of

Master of Science

Thesis: DATA-DRIVEN MODELING AND ANALYSIS FOR CARDIOVASCULAR DISEASE RISK PREDICTION AND REDUCTION

Major Field: Industrial Engineering & Management

Biographical:

Education:

Completed the requirements for the Master of Science in Industrial Engineering and Management Department at Oklahoma State University, Stillwater, OK in May 2021.

Completed the requirements for the requirements for the Bachelor's of Science in Industrial Engineering, and a Bachelor's of Arts in Economics at Antalya Bilim University and graduated with a double major in May 2019.

Experience:

Graduate Teaching Assistant-School of Industrial Engineering & Management, Oklahoma State University (August 2019 – May 2021).