

MACHINE LEARNING METHODS AND SYSTEMS FOR PROFILING RISK FOR CORONARY ARTERY DISEASE

TECHNICAL FIELD

5

The present disclosure relates to machine learning and in particular to the use of machine learning in the profiling of risk for coronary artery disease in a patient from cardiac data.

10 BACKGROUND

Coronary Artery Disease (CAD) is the leading single cause of death in the world, accounting for over 9 million deaths globally in 2016. By 2035, the incidence of CAD is estimated to rise by more than 40%. Owing to growing affluence and sedentary
15 lifestyles, people living in developed countries are particularly at risk. The Society of Actuaries reports that heart diseases are number one cause of preventable deaths in the United States from 2010 to 2015.

At the same time, emerging wearable technology and artificial intelligence-based
20 predictions for heart disease are being developed to disrupt traditional means of diagnosis and could prove effective in our fight against this latent disease. In collaboration with Stanford Medicine, Rock Health conducted a survey over 5 years to understand digital health adoption rates in every day consumers. From wearable ownership to digital health tracking, the survey reported an at least 2-fold increase in
25 every aspect of adoption rate from 2015 to 2019.

SUMMARY

According to a first aspect of the present disclosure, a machine learning method of risk
30 profiling for coronary artery disease in a patient is provided. The method comprises: receiving cardiac data for the patient; processing the cardiac data to extract a plurality of heart rate variability metrics for the patient; determining a plurality of domain risk scores from the plurality of heart rate variability metrics for the patient; inputting a classifier input comprising the plurality of domain risk scores into a classifier to

determine a coronary artery disease risk indication for the patient; and outputting an indication of the coronary artery disease risk for the patient.

Embodiments of the method use constituent risk scores to predict a subject's clinical status. If a subject is predicted to have CAD, then the total risk score quantifies how severe disease might be.

In an embodiment, wherein the classifier input further comprises an age / gender risk score. The method may further comprise looking up the age / gender risk score in a look up table using an age and a gender for the patient. Embodiments of the method can be applied across the population (different age groups and gender) without compromising on predictive value – which is reduced by age and gender confounders. Comparing between the age-gender adjusted risk score and an unadjusted age risk score, it has been found that the age-gender adjusted risk score performed better.

Age and gender are sensitive confounding factors when predicting CAD using HRV metrics. We adjusted the age-gender risk score based on empirical evidence that older people are more susceptible to CAD and that males are more predisposed to it.

In an embodiment, the plurality of heart rate variability metrics are each associated with one of a plurality of domains and each domain risk score of the plurality of domain risk scores is determined from the heart rate variability metrics associated with the respective domain. Individual constituent risk scores tell trained clinicians which domain contributes the most to total risk score. For example, if frequency domain risk score is much higher than the other risk scores, it could mean that the subject may not have been lying still during the screening/measurement process.

The plurality of domains may comprise one or more of the following: indexes, time domain, frequency domain and non-linear metrics.

In an embodiment, the method further comprises pre-processing the cardiac data for the patient to identify a continuous period of uncorrupted cardiac data for the patient. The continuous period of uncorrupted cardiac data for the patient is at least 5 minutes in duration. The cardiac data may be electrocardiography (ECG) data or

photoplethysmogram (PPG) data. Embodiments of the method comprise a data pre-processing algorithm that identifies an uncorrupted 5 minute segment of ECG or PPG recording and extracts relevant HRV metrics from it. The HRV metrics are then used to train the AI/ML classifier.

5

The classifier may implement at least one of a support vector machine (SVM) algorithm, a k-nearest neighbour algorithm, a linear regression algorithm, a random forest algorithm, an adaptive boost algorithm, and an artificial neural network algorithm such as a multi-layer perceptron algorithms.

10

The indication of the coronary artery disease risk for the patient may comprise a risk score.

For long-term monitoring, the CAD risk monitoring may be integrated with sleep stage analysis to ensure that we can acquire artefact-free ECG or PPG data. The CAD risk score does not need to be calculated continuously and can be performed at least once a day during sleep.

A smart device may be used to capture the patient cardiac data. Constant and remote monitoring is made possible throughout the day. Embodiments may leverage on the hardware of cutting edge smart devices to compute the risk score.

According to a second aspect of the present disclosure, a training method for a machine learning system for risk profiling for coronary artery disease in a patient is provided. The training method comprises: receiving training data comprising patient cardiac data, and patient coronary artery disease diagnosis data; extracting a plurality of heart rate variability metrics from the patient cardiac data; and training a classifier to model a probability that the patient coronary artery disease diagnosis data indicates disease as a function of features of heart rate variability metrics.

30

In an embodiment, the method further comprises normalizing each metric on a scale to determine a plurality of scaled metrics.

In an embodiment, the heart rate variability metrics correspond to a plurality of domains and wherein training the classifier comprises determining a sum of scaled metrics for each domain of the plurality of domains and training the classifier using the sum of scaled metrics for each domain.

5

In an embodiment, the training data further comprises patient age / gender data and training the classifier comprises taking into account an age / gender score.

10 According to a third aspect of the present disclosure a computer readable medium storing processor executable instructions which when executed on a processor cause the processor to carry out a method as set out above is provided.

15 The algorithm can be offered as a cloud application and it processes user data from a cloud server. The cloud machine learning service allows us to use consented data to improve our algorithm and make more accurate predictions, benefitting all users of the software. The algorithm can also be made available as a Software as a Service (SaaS) for purchase on any compatible wearable device that has a pulse oximeter, allowing users to access our service without having to purchase new hardware.

20 According to a fourth aspect of the present disclosure, a data processing system for risk profiling for coronary artery disease in a patient is provided. The system comprises: a processor, and a data storage device storing computer program instructions operable to cause the processor to: receive cardiac data for the patient; process the cardiac data to extract a plurality of heart rate variability metrics for the patient; determine a plurality of domain risk scores from the plurality of heart rate variability metrics for the patient; input a classifier input comprising the plurality of domain risk scores into a classifier to determine a coronary artery disease risk indication for the patient; and output an indication of the coronary artery disease risk for the patient.

30

The classifier input may further comprise an age / gender risk score. The data storage device may further store computer program instructions operable to cause the processor to: look up the age / gender risk score in a look up table using an age and a gender for the patient.

In an embodiment, the plurality of heart rate variability metrics are each associated with one of a plurality of domains and each domain risk score of the plurality of domain risk scores is determined from the heart rate variability metrics associated with the
5 respective domain.

In an embodiment, the data storage device further stores computer program instructions operable to cause the processor to: pre-process the cardiac data for the patient to identify a continuous period of uncorrupted cardiac data for the patient.

10 BRIEF DESCRIPTION OF THE DRAWINGS

In the following, embodiments of the present invention will be described as non-limiting examples with reference to the accompanying drawings in which:

15 FIG.1 is a block diagram showing a coronary artery disease risk profiling system according to an embodiment of the present invention;

FIG.2 is a flow chart showing a method of risk profiling for coronary artery disease in
20 a patient according to an embodiment of the present invention;

FIG.3 shows a method of training a coronary artery disease risk profiling system according to an embodiment of the present invention;

25 FIG.4 shows a scatter plot of probability vs sum of scaled metrics for time domain used in training method of an embodiment of the present invention;

FIG.5 shows the assignment of domain specific risk score based on the sum of scaled metrics and probability divisions in a training method of an embodiment of the present
30 invention;

FIG.6 shows the annual number of adults diagnosed with heart disease by gender and age group which is taken from a study by L. Mosca, E. Barrett-Connor, and N. K. Wenger, "Sex/Gender Differences in Cardiovascular Disease Prevention What a

Difference a Decade Makes,” *Circulation*, vol. 124, no. 19, pp. 2145–2154, Nov. 2011, doi: 10.1161/CIRCULATIONAHA.110.968792;

FIG.7 is a histogram showing the distribution of total risk score across clinical status
5 in a study cohort; and

FIG.8 is a block diagram showing an overview of evaluation methods of risk profiling
for coronary artery disease.

10 DETAILED DESCRIPTION

The present disclosure aims to capitalize on global trends and the potential for digital health to alleviate worrying numbers of CAD cases. An AI/ML algorithm to predict and quantify the severity of hypertension and CAD using only capabilities afforded by a
15 wearable health monitor is proposed. We use Heart Rate Variability metrics extracted from an ECG or PPG of a wearable device may be to make predictions. In embodiments, the algorithm requires only a short recording of 5-minutes. The ease of use makes it easy to be adopted by clinicians or every day users, and may also be deployed as a screening tool to identify high risk asymptomatic CAD patients.

20 Heart Rate Variability (HRV) metrics are used to measure autonomic nervous system (ANS) function is a widely researched field. ANS function is sensitive to factors such as age or severity of CAD, and strong empirical evidence demonstrate the correlation of HRV metrics with both factors. It is thought that the sinoatrial (SA) node of the heart
25 directly influences HRV, and a dysfunctional SA node – exacerbated by CAD, may be identified by discerning patterns and fluctuations in HRV.

The present disclosure relates to the provision and training an artificial intelligence (AI) / machine learning (ML) model with a unique combination of HRV metrics – to predict
30 if CAD is present in an individual and then proceeding to assign a numerical risk score to quantify its severity. To demonstrate the efficacy of the model, five common performance measures are used: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

FIG.1 is a block diagram showing a coronary artery disease risk profiling system according to an embodiment of the present invention. The coronary artery disease risk profiling system 100 is a computer system with memory that stores computer program modules which implement methods for risk profiling for coronary artery disease in patients according to embodiments of the present invention.

The coronary artery disease risk profiling system 100 comprises a processor 110, a working memory 120, an input interface 130, an output interface 140, program storage 150 and data storage 170. The processor 110 may be implemented as one or more central processing unit (CPU) chips. The program storage 150 is a non-volatile storage device such as a hard disk drive which stores computer program modules. The computer program modules are loaded into the working memory 120 for execution by the processor 150. The data storage 170 is a non-volatile storage device such as a hard disk drive which stores machine learning models and data. The input interface 130 is an interface which allows data including patient cardiac data to be received by the coronary artery disease risk profiling system 100. The output interface 140 is coupled to an output device such as a display which allows the results of the risk profiling to be displayed and analyzed.

The program storage 150 stores a cardiac data pre-processing module 152, a metric extraction module 154, a domain risk score calculation module 156, a coronary artery disease risk calculation module 158, and a classifier training module 160. The computer program modules cause the processor 110 to execute data processing which is described in more detail below. The program storage 150 may be referred to in some contexts as computer readable storage media and/or non-transitory computer readable media. As depicted in FIG.1, the computer program modules are distinct modules which perform respective functions implemented by the coronary artery disease risk profiling system 100. It will be appreciated that the boundaries between these modules are exemplary only, and that alternative embodiments may merge modules or impose an alternative decomposition of functionality of modules. For example, the modules discussed herein may be decomposed into sub-modules to be executed as multiple computer processes, and, optionally, on multiple computers. Moreover, alternative embodiments may combine multiple instances of a particular module or sub-module. It will also be appreciated that, while a software

implementation of the computer program modules is described herein, these may alternatively be implemented as one or more hardware modules (such as field-programmable gate array(s) or application-specific integrated circuit(s)) comprising circuitry which implements equivalent functionality to that implemented in software.

5

Although the coronary artery disease risk profiling system 100 is described with reference to a computer, it should be appreciated that the coronary artery disease risk profiling system 100 may be formed by two or more computers in communication with each other that collaborate to perform a task. For example, but not by way of limitation,
10 an application may be partitioned in such a way as to permit concurrent and/or parallel processing of the instructions of the application. Alternatively, the data processed by the application may be partitioned in such a way as to permit concurrent and/or parallel processing of different portions of a data set by the two or more computers. In an embodiment, virtualization software may be employed by the coronary artery disease
15 risk profiling system 100 to provide the functionality of a number of servers that is not directly bound to the number of computers in the coronary artery disease risk profiling system 100. In an embodiment, the functionality disclosed above may be provided by executing the application and/or applications in a cloud computing environment. Cloud computing may comprise providing computing services via a network connection using
20 dynamically scalable computing resources. A cloud computing environment may be established by an enterprise and/or may be hired on an as-needed basis from a third party provider.

The data storage 170 stores a classifier 172 and age gender risk scores 174. The
25 classifier 172 is a predictive model which is trained to take domain risk scores as input and to generate a coronary artery disease risk score for the patient based on the domain risk scores. The classifier 172 may be implemented using artificial intelligence or machine learning algorithms. Various types of artificial intelligence or machine learning algorithms may be used: classical machine learning, ensemble methods and
30 a supervised neural network. In classical machine learning, two distance-based algorithms: Support Vector Machine (SVM) and k-Nearest Neighbour (KNN), and one linear model: Logistic Regression may be adopted. For ensemble methods, the Random Forest (RF) and Adaptive Boost (AB) algorithms may be used. Finally, a Multi-layer perceptron (MLP) artificial neural network may be used. The input features

for all classifiers may include age gender risk score, index domain risk score, time domain risk score, frequency domain risk score and non-linear domain risk score. The classifications are either 'healthy' or 'with disease'. The gender scaling factors 174 are scaling factors which are created to account for gender bias when predicting CAD.

5

FIG.2 is a flow chart showing a method of risk profiling for coronary artery disease in a patient according to an embodiment of the present invention. The method 200 shown in FIG.2 is carried out by the coronary artery disease risk profiling system 100 shown in FIG.1.

10

In step 202, the input interface 130 of the coronary artery disease risk profiling system 100 receives cardiac data for the patient. The cardiac data may comprise Photoplethysmography (PPG) and / or Electrocardiography (ECG) signals collected for the patient. The cardiac data may be collected by a wearable health monitor or smart watch device. The cardiac data comprises of a signal collected over a time period of, for example, at least 5 minutes. As mentioned below, the received cardiac data may be pre-processed to identify an uncorrupted period of cardiac data, thus the input cardiac data may comprise a sequence of data relating to a longer time period from which a suitable portion is identified for use in the later processing.

20

In step 204, the cardiac data pre-processing module 154 is executed by the processor 110 of the coronary artery disease risk profiling system 100 to pre-process the cardiac data. The pre-processing comprises identifying a continuous uncorrupted period of cardiac data. The continuous uninterrupted period of cardiac data may have a duration of at least 5 minutes.

25

In step 206, the metric extraction module 154 is executed by the processor 110 of the coronary artery disease risk profiling system 100 to extract heart rate variability metrics from the cardiac data. 27 unique heart rate variability (HRV) metrics have been identified various combinations of the metrics may be extracted in step 206. These metrics correspond to different domains such as indexes, time domain, frequency domain and non-linear metrics. Examples of the HRV metrics are as follows: parasympathetic nervous system (PNS) index, sympathetic nervous system (SNS) index, stress Index, mean R-R interval, mean heart rate (HR), standard deviation of

30

NN intervals (SDNN), root mean square of successive differences between normal heartbeats (RMSSD), mean number of times an hour in which the change in successive normal sinus (NN) intervals exceeds 50 ms (NN50), the proportion of NN50 divided by total number of NNs (pNN50), HRV Triangular Index, Baseline width of the RR interval histogram (TINN), heart rate variability (HRV) very low frequency (VLF) or low frequency (LF) or high frequency (HF) power, ratios of HRV power, percentage of HRV power, normalized units (nu) of HRV power, statistical metrics such as Poincare standard deviations (SD), approximate entropy (ApEn), sample entropy (SampEn), and detrended fluctuation analysis (DFA) alpha measures.

Table 1 below shows heart rate variability (HRV) metrics in different domains:

HRV Metrics			
Indexes	Time Domain	Frequency Domain	Non-linear
- PNS Index	- Mean RR	- HRV VLF Power ms ²	- Poincare SD1
- SNS Index	- Mean HR	- HRV LF Power ms ²	- Poincare SD2
- Stress Index	- SDNN	- HRV HF Power ms ²	- Poincare
	- RMSSD	- HRV LF/HF Power ms ²	SD2/SD1 ratio
	- NN50	ratio	- ApEn
	- pNN50	- HRV VLF Power %	- SampEn
	- HRV Triangular Index	- HRV LF Power %	- DFA, alpha 1
	- TINN	- HRV HF Power %	- DFA, alpha 2
		- HRV LF Power nu	
		- HRV HF Power nu	

In step 208, the domain risk score calculation module 156 is executed by the processor 110 of the coronary artery disease risk profiling system to calculate domain risk scores from the heart rate variability metrics. A domain risk score is calculated for each for each of a plurality of domains. The domain risk scores may be calculated for each of the domains shown in table 1. A sigmoid function which is split into a number of divisions may be used for each domain to determine a domain risk score from a sum of scaled features calculated from the HRV metrics within that domain.

In step 210, the coronary artery disease risk score calculation module 158 is executed by the processor 110 of the coronary artery disease risk profiling system 100 to look up an age and gender scaling factor for the patient. The patient's age and gender are used to determine an age gender risk score from the age gender risk scores 174 stored in the data storage 170 of the coronary artery disease risk profiling system 100.

In step 212, the coronary artery disease risk score calculation module 158 is executed by the processor 110 of the coronary artery disease risk profiling system 100 to determine a coronary artery risk score for the patient. The domain risk scores calculated in step 208 and the age gender risk score are input into the classifier 172 and the classifier 172 generates a coronary artery disease risk score for the patient.

In step 214, the output interface 140 of the coronary artery disease risk profiling system 100 outputs an indication of the coronary artery disease risk score for the patient.

The data collection and training for the coronary artery disease risk profiling system will now be described with reference to FIG.3.

FIG.3 shows a method of training a coronary artery disease risk profiling system according to an embodiment of the present invention. The method 300 shown in FIG.3 may be executed on the coronary artery disease risk profiling system 100 shown in FIG.1 to train the classifier 172 stored in the data storage 170.

In step 302, the input module 130 of the coronary artery disease risk profiling system 100 receives training data. The training data comprises ECG/PPG signals. In this example, the ECG/PPG signals have a recording length of at least 5 minutes, available clinical diagnosis (i.e. Healthy, hypertension (HTN) or coronary artery disease (CAD)), available age and gender information of each subject. The following datasets were used to train an exemplary embodiment: The Vortal dataset (Charlton P.H. and Bonnici T.B. et al. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram, Physiological Measurement, 37(4), pp. 610-26, 2016. DOI: 10.1088/0967-3334/37/4/610) contains healthy subjects, the SHAREE dataset (P. Melillo, R. Izzo, A. Orrico, P. Scala, M. Attanasio, M. Mirra, N. De Luca, L. Pecchia "Automatic prediction of cardiovascular and cerebrovascular

events using Heart Rate Variability analysis", Plos One, March 20, 2015. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118504>) contains hypertensive subjects and the MIMIC Critical Care database (MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: <http://www.nature.com/articles/sdata20163515>) contains CAD subjects.

In step 304, the cardiac data pre-processing module 152 is executed by the processor 110 of the coronary artery disease risk profiling system 100 to pre-process the training data to identify continuous period of uncorrupted cardiac data. For each subject in each dataset, signal filtering and processing techniques are applied to extract 5 minutes of uncorrupted PPG/ECG signal. A segment was considered uncorrupted if all peaks were correctly identified with our peak detection algorithm in MATLAB that is based on signal segmentation using a moving window method. Using the datasets mentioned above, a total of 184 subjects met these criteria and they form our study cohort. Table 2 shows a summary of the study cohort. Table 3 shows how the cohort is split into training and test set for the AI/ML model training and evaluation.

Table 2 below shows Cohort Statistics

Group	Diagnosis	Number of Subjects	Number of Male Subjects	Age Range	Source
1	Healthy	49	25	18 - 79	VORTAL Dataset
2	Hypertension	75	50	46 - 86	SHAREE Dataset
3	CAD	60	42	47 - 79	MIMIC Critical Care Database

Table 3 below shows Cohort Allocation for Training and Test split

Group	Number of Subjects	Percentage (%)
-------	--------------------	----------------

Training	147	80
Test	37	20
Total	184	100

Steps 306-322 are carried out by the classifier training module executed by the processor 110 of the coronary artery disease risk profiling system 100.

- 5 In step 306, datasets corresponding to different diagnoses are identified. To create a binary classification for the logistic regression, HTN and CAD subjects to the were assigned to the same group and treated this group as 'with disease'. The healthy population was treated as 'healthy'.
- 10 In step 308, heart rate variability metrics are extracted and categorized into domains. Kubios HRV Standard Version 3.3.1 was used to extract the values of the 27 HRV metrics (Table 1) from each signal.

Table 4 below shows HRV time domain metrics in binary class the status 0 represents 'healthy' subjects and 1 represents subjects 'with disease'.

Subject ID	Status	Mean RR	SDNN	Mean HR	RMSSD	pNN50	TINN
52	0	917.38	62.196	65.403	42.367	20.859	288
53	0	1069.7	52.588	56.093	52.867	37.634	291
56	0	923.09	50.415	64.999	41.176	15.48	245
57	0	856.85	35.978	70.024	23.616	3.1519	171
h1911	1	654.37	14.915	91.691	9.6722	0	62
h2012	1	844.46	39.527	71.052	26.441	4.5198	204
h2020	1	1007.1	41.046	59.577	52.937	21.622	203
h2032	1	952.19	29.816	63.013	24.988	1.5924	138
h2035	1	1062.1	28.375	56.489	23.16	3.2028	135
h2041	1	1067.4	26.675	56.211	17.625	1.7857	124
h2047	1	1023.3	41.874	58.633	36.103	9.589	215
h2062	1	1002.9	39.642	59.825	32.619	4.3624	233

h2065	1	948.58	19.914	63.253	21.965	1.5873	101
-------	---	--------	--------	--------	--------	--------	-----

In step 310, each type of heart rate variability metric is normalized/scaled from a scale of 0 to 1.

- 5 Table 5 below shows normalized (scaled) HRV time domain metrics.

Subject ID	Status	Mean RR	SDNN	Mean HR	RMSSD	pNN50	TINN
52	0	0.582793	0.318883	0.310759	0.594023	0.647557	0.363196
53	0	0.434376	0.433672	0.198668	0.472918	0.364119	0.355932
56	0	0.577229	0.459633	0.305895	0.607759	0.738443	0.467312
57	0	0.641771	0.632115	0.366395	0.810292	0.946744	0.646489
h1911	1	0.839053	0.88376	0.62726	0.971116	1	0.910412
h2012	1	0.653844	0.589715	0.378771	0.777709	0.923631	0.566586
h2020	1	0.495372	0.571567	0.240615	0.472111	0.634665	0.569007
h2032	1	0.548875	0.705734	0.281984	0.794467	0.973094	0.726392
h2035	1	0.441781	0.72295	0.203436	0.815551	0.945884	0.733656
h2041	1	0.436617	0.74326	0.200089	0.87939	0.969828	0.760291
h2047	1	0.479587	0.561675	0.229249	0.66627	0.83798	0.539952
h2062	1	0.499464	0.588341	0.243601	0.706453	0.926291	0.496368
h2065	1	0.552392	0.824035	0.284873	0.829334	0.97318	0.815981

- In step 312, a binary class logistic regression is performed for each domain. The logistic regression gives the probability that a subject is 'with disease' as a function of the scaled HRV metrics belonging to that domain. For example, 6 metrics – Mean RR, SDNN, Mean HR, RMSSD, pNN50, TINN belong to the time domain and they form the logit function in the logistic regression.

- In step 314, a scatter graph plot of probability of adverse conditions vs sum of normalized domain metrics of each subject is generated. Each subject has a unique probability to whether they have the disease. They also have unique values of each scaled metric. We sum the values of the scaled metrics to obtain a single value, called

the sum of scaled metrics. This allows us to form a scatter plot – a graph of probability vs sum of scaled metrics for each of the 184 subjects in our study cohort.

FIG.4 shows a scatter plot of probability vs sum of scaled metrics for time domain used in a training method of an embodiment of the present invention. As shown in FIG.4, the probability of disease is plotted against the sum of scaled features for each subject.

In step 316, the scatter plots are interpolated using a LOWESS function (locally weighted scatterplot smoothing). This generates a single sigmoid function.

The interpolated sigmoid function 410 for the time domain is shown in FIG.4.

In step 318, the probability of adverse conditions is split into divisions. The probability of disease can be split into k divisions, where k represents a fixed number. k = 10 is chosen for this study. Splitting the probability scale into 10 divisions creates 10 risk scores for every domain. Each risk score corresponds to an abscissa range (i.e. the sum of scaled domain metrics) based on the sigmoid function.

FIG.5 shows the assignment of domain specific risk score based on the sum of scaled metrics and probability divisions in a training method of an embodiment of the present invention.

As shown in FIG.5, the probability of disease scale is split into 10 divisions: a domain risk score 1 512 corresponds to a probability of disease from 0.0 to 0.1, a domain risk score 9 514 corresponds to a probability of disease from 0.8 to 0.9 and a domain risk score 10 516 corresponds to a probability of disease from 0.9 to 1.0. The domain risk score 1 512 corresponds to a sum of scaled features the range 0 to 2.5. The domain risk score 10 516 corresponds to a sum of scaled features in the range 5.2 to 5.

In step 320, age / gender risk scores are created. Several studies demonstrate the relationship between gender and the incidence of heart disease. In many findings, men are reported to be more predisposed to developing heart disease than women across all age groups. There have been many explanations including differences in estrogen levels that affect lipid metabolism and LDL cholesterol levels. The age

adjusted risk score takes into account gender biases by introducing a gender scaling factor into the original age risk score. We derive our gender scaling factor from a 2011 study (L. Mosca, E. Barrett-Connor, and N. K. Wenger, “Sex/Gender Differences in Cardiovascular Disease Prevention What a Difference a Decade Makes,” *Circulation*, vol. 124, no. 19, pp. 2145–2154, Nov. 2011, doi: 10.1161/CIRCULATIONAHA.110.968792) on gender differences in cardiovascular disease prevention.

FIG.6 shows the annual number of adults diagnosed with heart disease by gender and age group which is taken from a study by L. Mosca, E. Barrett-Connor, and N. K. Wenger, “Sex/Gender Differences in Cardiovascular Disease Prevention What a Difference a Decade Makes,” *Circulation*, vol. 124, no. 19, pp. 2145–2154, Nov. 2011, doi: 10.1161/CIRCULATIONAHA.110.968792. It is clear from FIG.6 that the likelihood of men developing the disease is higher than that of women. Based on this, Thus, a gender scaling factor can be calculated as shown in table 6 below.

Age Group	Gender Scaling Factor
< 45	2.40 (extrapolated)
45 – 54	2.40
55 – 64	2.40
65 – 74	1.73
75 – 84	1.47
85 – 94	1.14

Based on the gender scaling factor and the age risk score, an age / gender risk score can be calculated for each subject. These are shown in Table 7 below.

Subject ID	Gender Scaling Factor	Age Risk Score	Adjusted Age Risk Score
52	1	2	2
53	2.4	1	2.4

56	1	1	1
57	1	1	1
h1911	2.4	6	14.4
h2012	1.73	9	15.57
h2020	1.47	10	14.7
h2032	1	8	8
h2035	1.47	10	14.7
h2041	1.14	10	11.4
h2047	1	8	8
h2062	1.73	9	15.57
h2065	1	8	8

Finally, each subject will have a total of 5 domain risk scores – Age adjusted, index, time, frequency and non-linear. These are shown in Table 8 below.

Subject ID	Status	Index Domain	Time Domain	Frequency Domain	Non-linear Domain	Adjusted Age Risk Score	Total Risk Score
52	0	3	2	1	2	2	10
53	0	2	1	2	2	2.4	9.4
56	0	3	3	2	6	1	15
57	0	5	7	4	4	1	21
h1911	1	10	10	10	9	14.4	53.4
h2012	1	5	6	4	3	15.57	33.57
h2020	1	2	3	4	6	14.7	29.7
h2032	1	4	7	6	7	8	32
h2035	1	5	6	9	8	14.7	42.7
h2041	1	5	7	7	8	11.4	38.4
h2047	1	3	4	3	4	8	22
h2062	1	3	4	5	4	15.57	31.57
h2065	1	6	8	10	8	8	40

FIG.7 is a histogram showing the distribution of total risk score across clinical status in the study cohort. As can be seen in FIG.7, the total risk score differentiates between healthy patients and patients with disease. Additionally, patients with coronary artery disease (CAD) generally have a higher total risk score than those with hypertension (HTN).

In step 322, the risk scores are used to train the classifier. The input features for the classifier are age adjusted risk score, index domain risk score, time domain risk score, frequency domain risk score and non-linear domain risk score. The classifications are either 'healthy' or 'with disease' (combination of HTN and CAD subjects).

The performance of systems and methods according to embodiments of the present invention was evaluated. 5 performance measurements were used to evaluate each model: Accuracy, Sensitivity, Specificity, PPV and NPV. The models were evaluated at two instances: a hold-out approach (20% of observations were used for the independent test set) and a 5-fold nested cross validation (80% of observations were used for the training set).

FIG.8 is a block diagram showing an overview of evaluation methods of risk profiling for coronary artery disease.

As shown in FIG.8, the input to the evaluation is the set of domain risk scores for each of a set of subjects: the age adjusted score 810, the index domain score 812, the time domain score 814, the frequency domain score 816, and the non-linear domain score 818. The data is divided into a training set 820 and a test set 830. As mentioned above, 80% of the data was used for the training set 820 and the remaining 20% was used for the test set 830. Nested k-fold cross validation 840 was carried out on the training set 820 and hold out evaluation 850 was carried out using the test set 830.

The training data set 820 is divided into a training inner subset 822, a validation inner subset 824, a training outer subset 826 and a validation outer subset 828. The training outer subset 826 and the validation outer subset 828 are used in model selection 842. Following model selection 842, the training inner subset 822 and the validation inner subset 824 are used for hyperparameter tuning 844 on the selected model.

k-fold performance metrics 846 are calculated and the best hyperparameters 848 are determined. The best hyperparameters 848 are then used with the model to perform hold out evaluation 850 on the test set 830. This gives a set of hold-out performance metrics 852.

Artificial neural network and distance-based machine learning algorithms were the best performing models in terms of accuracy. This was followed by Logistic Regression and Ensemble methods. Ensemble methods were the poorest performing models in terms of accuracy.

In the cross validation, MLP, SVM and KNN had the highest accuracies ranging from 90.5% to 91.2%. Similarly, they were the best performing in sensitivity ranging from 96.3% to 99.1% and NPV ranging from 88.2% to 96.4%. Conversely, RF and AB had the highest specificities ranging from 87.2% to 94.9% and PPV ranging from 94.7% to 97.3%.

In the hold out evaluation, MLP had the highest accuracy, sensitivity and NPV at 97.3%, 100% and 100% respectively. KNN had an accuracy, sensitivity and NPV of 94.6%, 100% and 100% respectively. AB and RF were the best performing at specificity and PPV, with both at 100%. However, their performance for accuracy ranged 70.3% to 78.4%, sensitivity ranged 59.3% to 70.4% and NPV ranged 47.6% to 55.6%.

Table 9 below shows performance measures for 6 AI/ML models via cross validation

5-fold Cross Validation Performance

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
MLP	90.5	98.1	69.2	89.8	93.1
SVM	91.2	96.3	76.9	92.0	88.2
KNN	91.2	99.1	69.2	89.9	96.4
LR	87.1	90.7	76.9	91.6	75.0
AB	83.7	82.4	87.2	94.7	64.2
RF	73.5	65.7	94.9	97.3	50.0

Table 10 below shows Performance measures for 6 AI/ML models via hold out evaluation.

Test Set Performance (Hold-out evaluation)

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
MLP	97.3	100.0	90.0	96.4	100.0
SVM	94.6	96.3	90.0	96.3	90.0
KNN	94.6	100.0	80.0	93.1	100.0
LR	86.5	85.2	90.0	95.8	69.2
AB	78.4	70.4	100.0	100.0	55.6
RF	70.3	59.3	100.0	100.0	47.6

While the AI/ML models are binary classifiers for our study cohort i.e. ‘Healthy’ vs ‘Diseased’, our total risk score assignment can be used to further distinguish between the original clinical statuses (i.e. Healthy vs HTN vs CAD) – as can be seen from FIG.7.

- 5 For this study, we demonstrate the feasibility of this method. The total risk score threshold for each clinical status is based on achieving the highest accuracy possible for that class.

Table 11 below shows the accuracy of classifications based on risk score threshold for each class

Total Risk Score (Range threshold).	Classified as	Actual observation	Accuracy
< 23	Healthy	Healthy = 34 HTN = 1 CAD = 0	94.4%
23 – 48	HTN	Healthy = 15 HTN = 64 CAD = 26	61.0%
> 48	CAD	Healthy = 0 HTN = 10 CAD = 34	77.3%

The methods and results described above demonstrates the feasibility of our predictive algorithm on 5 minutes of PPG or ECG recording. Briefly summarized, our applied AI/ML algorithm first predicts if a subject falls under the category of ‘Healthy’ or ‘Diseased’ – which consists of either HTN or CAD clinical diagnosis. Our cross

validated accuracy of the initial prediction is 91.2% using either SVM or KNN classifiers. Concurrently, a quantifiable total risk score is generated. Higher total risk scores are positively associated with severe HTN and CAD. By assigning total risk score thresholds, we are able to classify Healthy subjects with 94.4% accuracy, HTN subjects with 61.0% accuracy and CAD subjects with 77.3% accuracy.

The age-gender adjusted risk score was created to account for gender biases when predicting CAD. Several studies show that males are more predisposed to the disease and hence should be assigned a higher risk score. To demonstrate the improvement of the age-gender adjusted risk score over an unadjusted risk score, we performed a 5-fold cross validation using the 3 best performing models – MLP, SVM and KNN. In one instance, the age-gender adjusted risk score was used and the unadjusted risk score was used in another. All other inputs (domain risk scores) remain the same.

Table 12 below shows a comparison between age-gender adjusted risk score vs unadjusted risk score.

Model	Age-Gender adjusted risk score			Unadjusted risk score		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
MLP	90.5%	98.1%	69.2%	89.1%	95.4%	71.8%
SVM	91.2%	96.3%	76.9%	89.8%	95.4%	74.4%
KNN	91.2%	99.1%	69.2%	87.1%	94.4%	66.7%

Another important consideration to our study is the type of performance metric used to evaluate our model. Because of application deals with predicting HTN and CAD from an otherwise healthy population, we focus on eliminating false negatives – where a subject is predicted as healthy but in fact has the underlying disease. NPV is a good measure for false negatives and it is calculated as a percentage of true negatives divided by the sum of true negatives and false negatives. High NPV values indicate low incidence of false negatives. With our unique risk score features as inputs to the

classifier of the present disclosure, we are able to achieve a cross validated NPV of 96.4% on the KNN classifier.

Table 13 below shows a comparison of models of the present disclosure with meta-analysis of AI/ML algorithms in predicting CAD.

Performance metrics	Meta-analysis (<i>Krittanawong et al.</i>)			Models of the present disclosure		
	CNN	Boosting	Custom	MLP	SVM	KNN
AUC	0.87	0.88	0.93	0.95	0.93	0.90
Sensitivity	0.86	0.86	0.87	0.98	0.96	0.99
Specificity	0.70	0.70	0.86	0.69	0.77	0.69

Krittanawong et al. (C. Krittanawong et al., “Machine learning prediction in cardiovascular diseases: a meta-analysis,” Scientific Reports, vol. 10, no. 1, p. 16057, Sep. 2020, doi: 10.1038/s41598-020-72685-1.) reports almost 15,025 references to AI/ML in predicting cardiovascular diseases from 1966 to 2019. In the same study, a meta-analysis of AI/ML algorithms used to predict CAD was conducted. A total cohort size of 45 containing 116,227 individuals was used in developing various algorithms that was mainly classified into 3 types: Convolutional Neural Networks, Boosting algorithms and Custom-built AI/ML algorithms. A comparison of the pooled performance of each type of algorithm reported in the study was done with 3 of the best performing models of the present disclosure. Only AUC, Sensitivity and Specificity were measured in the meta-analysis.

Overall, the models of the present disclosure performed better in sensitivity, ranging from 0.96 to 0.99 compared to the meta-analysis range of 0.86 to 0.87. The pooled specificity in the meta-analysis ranged from 0.70 to 0.86. Our models had a specificity that ranged from 0.69 to 0.77. This indicates that the models of the present disclosure likely had higher false positives but had significantly lower false negatives in the disease.

The AI/ML algorithm applied in the present disclosure predicts HTN and CAD from healthy populations using only 5 minutes of ECG or PPG data, while providing a numerical risk score that makes quantifying the severity of HTN and CAD easier for clinicians or users of digital health tools. The total numerical risk score comprises of various HRV domain risk scores and an age-gender adjusted risk score. Each domain risk score is derived from a Logistic Regression LOWESS plot of probability vs the sum of scaled metrics belonging to each domain. Thus, the algorithm provides ease of use for HTN and CAD risk screening with its short screening duration and easy to interpret risk score.

Whilst the foregoing description has described exemplary embodiments, it will be understood by those skilled in the art that many variations of the embodiments can be made within the scope and spirit of the present invention.

CLAIMS

1. A machine learning method of risk profiling for coronary artery disease in a patient, the method comprising:
 - receiving cardiac data for the patient;
 - processing the cardiac data to extract a plurality of heart rate variability metrics for the patient;
 - determining a plurality of domain risk scores from the plurality of heart rate variability metrics for the patient;
 - inputting a classifier input comprising the plurality of domain risk scores into a classifier to determine a coronary artery disease risk indication for the patient; and
 - outputting an indication of the coronary artery disease risk for the patient.
2. A method according to claim 1, wherein the classifier input further comprises an age / gender risk score.
3. A method according claim 2, further comprising looking up the age / gender risk score in a look up table using an age and a gender for the patient.
4. A method according to any preceding claim, wherein the plurality of heart rate variability metrics are each associated with one of a plurality of domains and each domain risk score of the plurality of domain risk scores is determined from the heart rate variability metrics associated with the respective domain.
5. A method according to claim 4, wherein the plurality of domains comprises one or more of the following: indexes, time domain, frequency domain and non-linear metrics.
6. A method according to any preceding claim further comprising pre-processing the cardiac data for the patient to identify a continuous period of uncorrupted cardiac data for the patient.
7. A method according to claim 6, wherein the continuous period of uncorrupted cardiac data for the patient is at least 5 minutes in duration.

8. A method according to any preceding claim, wherein the cardiac data is electrocardiography (ECG) data or photoplethysmogram (PPG) data.
9. A method according to any preceding claim, wherein the classifier implements at least one of a support vector machine (SVM) algorithm, a k-nearest neighbor algorithm, a linear regression algorithm, a random forest algorithm, an adaptive boost algorithm, and an artificial neural network algorithm such as a multi-layer perceptron algorithms.
10. A method according to any preceding claim wherein the indication of the coronary artery disease risk for the patient comprises a risk score.
11. A training method for a machine learning system for risk profiling for coronary artery disease in a patient, the training method comprising:
 - receiving training data comprising patient cardiac data, and patient coronary artery disease diagnosis data;
 - extracting a plurality of heart rate variability metrics from the patient cardiac data; and
 - training a classifier to model a probability that the patient coronary artery disease diagnosis data indicates disease as a function of features of heart rate variability metrics.
12. A training method according to claim 11, further comprising normalizing each metric on a scale to determine a plurality of scaled metrics.
13. A training method according to claim 12, wherein the heart rate variability metrics correspond to a plurality of domains and wherein training the classifier comprises determining a sum of scaled metrics for each domain of the plurality of domains and training the classifier using the sum of scaled metrics for each domain.
14. A training method according to any one of claims 11 to 13, wherein the training data further comprises patient age / gender data and wherein training the classifier comprises taking into account an age / gender score.

15. A computer readable medium storing processor executable instructions which when executed on a processor cause the processor to carry out a method according to any one of claims 1 to 14.

16. A data processing system for risk profiling for coronary artery disease in a patient, the system comprising a processor, and a data storage device storing computer program instructions operable to cause the processor to:

receive cardiac data for the patient;

process the cardiac data to extract a plurality of heart rate variability metrics for the patient;

determine a plurality of domain risk scores from the plurality of heart rate variability metrics for the patient;

input a classifier input comprising the plurality of domain risk scores into a classifier to determine a coronary artery disease risk indication for the patient; and

output an indication of the coronary artery disease risk for the patient.

17. A data processing system according to claim 16, wherein the classifier input further comprises an age / gender risk score.

18. A data processing system according to claim 17, wherein the data storage device further stores computer program instructions operable to cause the processor to: look up the age / gender risk score in a look up table using an age and a gender for the patient.

19. A data processing system according to any one of claims 16 to 18, wherein the plurality of heart rate variability metrics are each associated with one of a plurality of domains and each domain risk score of the plurality of domain risk scores is determined from the heart rate variability metrics associated with the respective domain.

20. A data processing system according to any one of claims 16 to 19, wherein the data storage device further stores computer program instructions operable to cause the processor to: pre-process the cardiac data for the patient to identify a continuous period of uncorrupted cardiac data for the patient.