# Comparative Performance Analysis of Random Forests against AutoPrognosis for predicting Coronary Heart Disease Risk and Metabolic Syndrome: A Retrospective Cohort Study

*Paulina Genet Ngcayiya*[1*] *and Dr Pravesh Ranchod*[2]

[1]1School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, South Africa
[2]School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, South Africa

**Abstract.** Cardiovascular Disease (CVD) is the leading cause of mortality worldwide. Amongst them, Coronary Heart Disease (CHD) is the most common type of CVD. The consequences of the presence of CVD risk factors often manifest as Metabolic Syndrome (MetS). In this study, data from the Framingham Heart Study (FHS), consisting of 4240 records and 17 variables, was used to build two types of 10-year CHD risk prediction models based on Random Forests (RF) and AutoPrognosis. The Framingham Risk Score model (AUC-ROC: 0.633) was used as a baseline model for performance evaluation. Results showed that the RF model with optimized hyperparameters had the best performance (AUC-ROC: 0.728). Furthermore, a dataset of 7821 records and 77 variables from the National Health and Nutrition Examination Survey (NHANES) was used to assess the predictive performance of RF against AutoPrognosis for determining the presence of MetS. The RF model with optimized hyperparameters had the best performance (AUC-ROC: 0.851). The performance of RF against AutoPrognosis on different sample sizes of data, ranging from 100 to 4900, was tested. The RF model with optimized hyperparameters had the best overall performance, followed by AutoPrognosis with an ensemble pipeline, then AutoPrognosis with a single pipeline and finally the RF model with default hyperparameter values.

## 1 Introduction

Cardiovascular Disease (CVD) is a medical term used to refer to a collection of diseases that affect the heart and blood vessels. CVD accounts for approximately 32% of all deaths worldwide, making it the leading cause of death globally [16]. Over 75% of CVD deaths occur in developing countries [16]. Coronary Heart Disease (CHD) is the most common type of CVD [16]. This disease develops when the blood vessels which supply the heart with blood

* Corresponding author: paulinangcayiya@gmail.com

and oxygen begin to harden and become narrow because of a build-up of fatty substances known as cholesterol. This often results in the development of chest pains or arrhythmia and can even progress to the occurrence of a heart attack or heart failure. The major determinants of CHD are risk factors such as old age, consuming an unhealthy diet, inadequate physical activity, alcohol abuse, and tobacco use (note that most of these are behavioural risk factors, meaning that they can be addressed through changes in lifestyle choices). The consequences of these behavioural risk factors can manifest themselves as Metabolic Syndrome (MetS). MetS is a medical term used to refer to a collection of conditions that can increase a person's risk of developing a CVD and/or diabetes. MetS includes hypertension, high blood glucose levels, unhealthy cholesterol levels, overweight, and obesity. These factors are often included as some of the core variables used in models developed for CVD risk prediction.

Most of the existing CVD risk prediction models, such as the Framingham Risk Score (FRS) model, were created and validated on data from American and European populations [7]. These are populations in developed countries where the number and severity of CVD risk factors differ from those in developing countries. Furthermore, a great number of the currently existing CVD risk prediction models used in the clinical setting make use of multivariate regression methods. Generally, these multivariate regression models use a limited number of conventional risk factors and assume that the correlation between all such factors and the CVD outcome is linear. These characteristics can limit a model's predictive performance, especially for certain subgroups of the population. Models based on ensemble machine learning algorithms often have the ability to realize more complex patterns within large datasets and model nonlinear relations for interactions between predictor variables and the outcome. Supervised learning has been used previously to model CVD risk outcomes, with the most successful methods being Random Forests [3] and AutoPrognosis [1].

In this study, the main objectives were to determine the most significant risk factors related to CHD and to conduct a performance evaluation study to compare the predictive performance between the FRS model and two different types of 10-year CHD risk prediction models based on ensemble Machine Learning (ML) methods, namely Random Forest (RF) and AutoPrognosis. The above procedure was performed on a dataset, of a cohort of 4133 people, from the Framingham Heart Study (FHS). The dataset was obtained from the Kaggle [17] website. Furthermore, as a novel study, a dataset of a cohort of 6781 people, from the National Health and Nutrition Examination Survey (NHANES), was used to assess the predictive performance of AutoPrognosis for specifically determining the presence of Metabolic Syndrome. The dataset was acquired from the Data World [18] website. The performance of RF against AutoPrognosis on different sample sizes of data, ranging from 100 to 4900, was also tested.

## 2 Materials and Methods

## 2.1 Modelling CHD risk prediction

### *2.1.1 Sample Design and Population*

FHS is an epidemiologic study based on CVD [20]. The study began in 1948 in the town of Framingham in Massachusetts, USA. When the study began, 5209 males and females aged between 30 and 60 were recruited by FHS researchers. They were used for the first course of wide-ranging physical examinations and interviews that were carried out. Thereafter, the subjects continued to return every two to six years to participate in further physical exams

and laboratory tests, while also providing detailed information on their medical history, as a means of updating information for the study. As the years progressed, the FHS transformed into a multigenerational study by collecting data from the two generations (children and grandchildren) of the original cohort. Initially, FHS was based on a largely white American population. However, since then the study has extended to become more inclusive of more ethnically diverse populations by enrolling participants of Hispanic, Indian, African American, Asian, Native American and Pacific Islander descent.

### *2.1.2 Data Collection and Preparation*

The FHS is still ongoing to date. The dataset used in the study presented by this paper is from the FHS and is publicly available on the Kaggle website [17]. The dataset has 4240 records and 17 variables for each record. After a series of data cleaning procedures, the dataset included 3658 records and 16 variables for each record. Data cleaning included the standard procedures of removing duplicated records, removing/imputing missing values, removing invalid points and outliers, deleting obvious error messages and data standardisation. Each variable is a possible CHD risk factor. The variables can be categorised into three categories of risk factors: demographic, medical and behavioural risk factors. The target variable was the outcome for the 10-year risk of CHD. The list of variables and definitions are provided in [17].

### *2.1.3 Models Tested*

**Framingham Risk Score (FRS)** - The FRS is a well-established and commonly used model for CVD risk prediction [9]. The datasets used to develop FRS all come from the Framingham Heart Study. Multiple adaptions and revisions of the original FRS model have been made over the years for assessing the risk of an individual developing a specific type of CVD. However, there also exists an FRS model that can be used to predict the general risk of an individual developing any type of CVD or CVD event. This model has been proven to perform just as well as the separate disease-specific variations of FRS, to predict an individual component of CVD, such as CHD. Furthermore, it is the model which was chosen to be used in the study outlined by this paper.

The standard version of the model includes HDL cholesterol as one of the variables in its algorithm. However, this is a variable which was not provided in the dataset used for this study. Fortunately, FRS provides a variation of the standard model based on non-laboratory variables, which substitutes the lipid variables such as HDL cholesterol with Body Mass Index (BMI). Therefore, in this study, we used the predicting equations of the BMI-based FRS model for general CVD risk prediction which was published in [21]. These equations were composed of beta-coefficients and survival functions. The model has seven core variables: sex, age, diabetes status, smoking status, treatment for hypertension, systolic blood pressure and BMI. All these variables were present in our dataset.

**Random Forest (RF)** - The RF algorithm is based on supervised learning and can be used for both classification and regression problems. CHD risk prediction is a classification problem. The algorithm works by creating multiple decision trees which are trained independently on random subsets of the training dataset. Additionally, random feature selection is used to split each node within a decision tree. The final output generated by the algorithm is the modal of the predicted classes from the multiple decision trees [3].

For this study, the RF model was implemented using the Scikit-learn library in the Python programming language. Our dataset was split into training and test datasets which are randomly selected to form a 70-30% split, respectively. The model was trained by fitting the Random Forest Classifier with the training data and tuning the parameters (maximum tree depth and maximum features) for optimization via grid search. The hyperparameter (the number of decision trees in the forest) was determined by assessing the out-of-bag error rate. The test data set was used to procure the AUC-ROC value to indicate the performance of the model on unseen data.

**AutoPrognosis** - AutoPrognosis was proposed fairly recently in 2018 by [1], where the practicality and effectiveness of this method were established using nine major patient groups characterizing various aspects of cardiovascular patient care. AutoPrognosis is an automated machine learning framework that uses an advanced Bayesian optimization technique to automatically generate a prognostic model made up of a weighted ensemble of machine learning pipelines. Each pipeline is made up of a data imputation, feature processing, classification, and calibration algorithm of its own. For training the model, AutoPrognosis is set to conduct several iterations of the Bayesian Optimization procedure. In each iteration, a new machine learning is explored, and its hyper-parameters are tuned. In this study, the number of iterations was set to 10. In every iteration, 5-fold cross-validation was used to assess the performance of the pipeline being evaluated. In this study, two different versions of the AutoPrognosis model were created: one based on the seven core variables used in the FRS model, and the other based on all the variables provided in our dataset.

(The code for the state-of-the-art AutoPrognosis model used in this study can be found at [22]).

### *2.1.4 Performance Evaluation*

The FRS model was used as the baseline model for a comparative performance evaluation of the proposed models used in this study. The AUC-ROC value for the FRS, RF and AutoPrognosis models were calculated. These values were attained after testing was done on the test datasets. Once all the models were developed and tested, a comparative analysis of their performances based on their AUC-ROC values was done to see model had better predictive abilities for CHD risk prediction.

(In some instances, where model improvements were necessary, interpretation of the results led to the need to return to the previous phases of training and testing to adjust the models and their parameters.)

### *2.1.5 Variable Ranking*

Each variable used to create a classification model contributes differently to the predictions made by the model. To determine the relative significance of each variable to the CHD risk prediction outcome, a random forest model was fitted to the data with the patients' variables as inputs and the predictions made by the model (the model which performed better between the AutoPrognosis or Random Forest model) as the outputs. Based on that, variable importance scores were allocated for each variable based on feature permutation.

## 2.2 Modelling Metabolic Syndrome outcome prediction

### *2.2.1 Sample Design and Population*

The NHANES [19] is an initiative of studies created to assess the health and nutritional status of both people in their youth and adulthood in the USA. NHANES began in 1959 and is still ongoing. Data is collected through a survey which combines physical examinations and interviews. The survey is carried out on a sample of approximately 5000 people for every iteration of data collection once a year. The people of the sample reside in different counties of the USA and are chosen to be nationally representative of all age and ethnicity groups in the USA population. Interviews are conducted in participants' homes and include socioeconomic, demographic, dietary and other health-related questions about lifestyle choices. The physical examinations are conducted in mobile health centres and include dental, medical, and physiological measurements and laboratory tests. Workers of the study consist of multiple physicians, health and medical technicians, as well as health and dietary interviewers.

### *2.2.2 Data Collection and Preparation*

The dataset used in the latter part of the study presented by this paper is from the NHANES and is publicly available on the Data World [18] website. The dataset has 7821 records and 77 variables for each record. After a series of data cleaning procedures, the dataset included 6781 records and 24 variables for each record. Data cleaning included the standard procedures of removing duplicated records, removing/imputing missing values, removing invalid points and outliers, deleting obvious error messages and data standardisation. A person is diagnosed with MetS if they have three or more of the following conditions: hypertension, diabetes, dyslipidemia, and overweight/obesity. Although, our dataset did not come with variables for the dyslipidemia and overweight/obesity conditions we were still able to produce them using the HDL and triglyceride cholesterol and waist circumference variables, respectively. The variables can be categorised into three categories of risk factors: demographic, medical and behavioural risk factors. The target variable, indicative of the presence/absence of MetS, was produced by checking if each patient was diagnosable with MetS based on the criteria stated above and assigned them with a binary value. The list of variables and definitions are provided in [23].

### *2.2.3 Models Tested, Performance Evaluation, Variable Ranking*

Two types of models were built for predicting the presence of MetS. The first model was based on Random Forests, while the other was based on AutoPrognosis. Information for the descriptions of the models and the procedures followed during the training and testing phases can be found under the *Models Tested* sub-section in the *Modelling CHD risk prediction* section further above, as the exact same information applies to this half of the study presented by this paper. Likewise, information on how performance evaluation and variable ranking was performed can be found in the *Performance Evaluation* and *Variable Ranking* sub-sections above.

### *2.2.4 Models Comparing the performance of Random Forests against AutoPrognosis on different sample sizes of data*

In [1] the performance of AutoPrognosis was evaluated using datasets with at least 30 000 patient records. Sometimes attaining data that large and training models on it can be challenging and computationally expensive. Therefore, for the final phase of the study, using the MetS dataset, we trained, tested and compared the performance of Random Forests

against AutoPrognosis on 13 different sample sizes of data ranging from 100 to 4900 (at intervals of 400). Models were trained and tested as described in sections further above.

# 3 Results

## 3.1 CHD Risk Prediction

### *3.1.1 Characteristics of the study population*

The dataset used in the study presented by this paper is from the Framingham Heart Study and is publicly available on the Kaggle website. A total number of 3 658 participants had enough data for inclusion in our study. The mean (std) age of participants was 49.55 (8.56) years, while 1 623 (44.37%) were male and 2 035 (55.63%) were female. 557 participants (15.23%) were predicted to have a 10-year risk of developed CHD, with the mean (std) age being 54.27 (7.99) where 307 (55.12%) are male and 250 (44.88%) are female.

### *3.1.2 Comparison of prediction models*

**Table 3.** Performance of 10-year CHD risk prediction models.

| Model | AUC-ROC | AUC Change |
|---|---|---|
| Framingham Risk Score | 0.633 | Baseline model |
| Random Forest (ignoring hyperparameters) | 0.653 | +2.0% |
| Random Forest (optimised hyperparameters) | 0.728 | +9.5% |
| AutoPrognosis (single pipeline) – using the 7 core variables | 0.703 | +7.0% |
| AutoPrognosis (ensemble pipeline) – using the 7 core variables | 0.696 | +6.3% |
| AutoPrognosis (single pipeline) – using all variables | 0.704 | +7.1% |
| AutoPrognosis (ensemble pipeline) – using all variables | 0.714 | +8.1% |

The AUC-ROC values for the different models under evaluation for 10-year CHD risk prediction are shown in Table 3. The Framingham Risk Score model (AUC-ROC: 0.633) was used as the baseline model for performance evaluation. There were two RF models under evaluation. For one of the RF models, hyperparameters were ignored during training (meaning that the hyperparameters were set to default values). For the other RF model, the hyperparameters were optimized using the out-of-bag error value. Both models outperformed the baseline model however, the RF model with optimized hyperparameters (AUC-ROC: 0.728) performed significantly better than the RF model (ignoring hypermeters) (AUC-ROC: 0.653). The AutoPrognosis model can be created with a single pipeline or an ensemble

pipeline. In some instances, the ensemble helps (however it was not always necessary). AutoPrognosis (single pipeline, using the 7 core variables) (AUC-ROC: 0.703) slightly outperformed AutoPrognosis (ensemble pipeline, using the 7 core variables) (AUC-ROC: 0.696). However, AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.714) outperformed AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.704). All these different versions of AutoPrognosis significantly outperformed the baseline model. The RF model with optimized hyperparameters had the best overall performance.

### *3.1.3 Variable Importance*

Figure 1 displays a list of all variables ranked based on their contribution to the RF model with optimized hyperparameters (best performing model). The importance scores are based on their mean decrease in impurity. Along with the conventional CHD risk factors, 'heartRate' was among the top ranking. Although the 'currentSmoker' and 'diabetes' variables were ranked the lowest, 'glucose' and 'cigsPerDay' were in the top ranking.

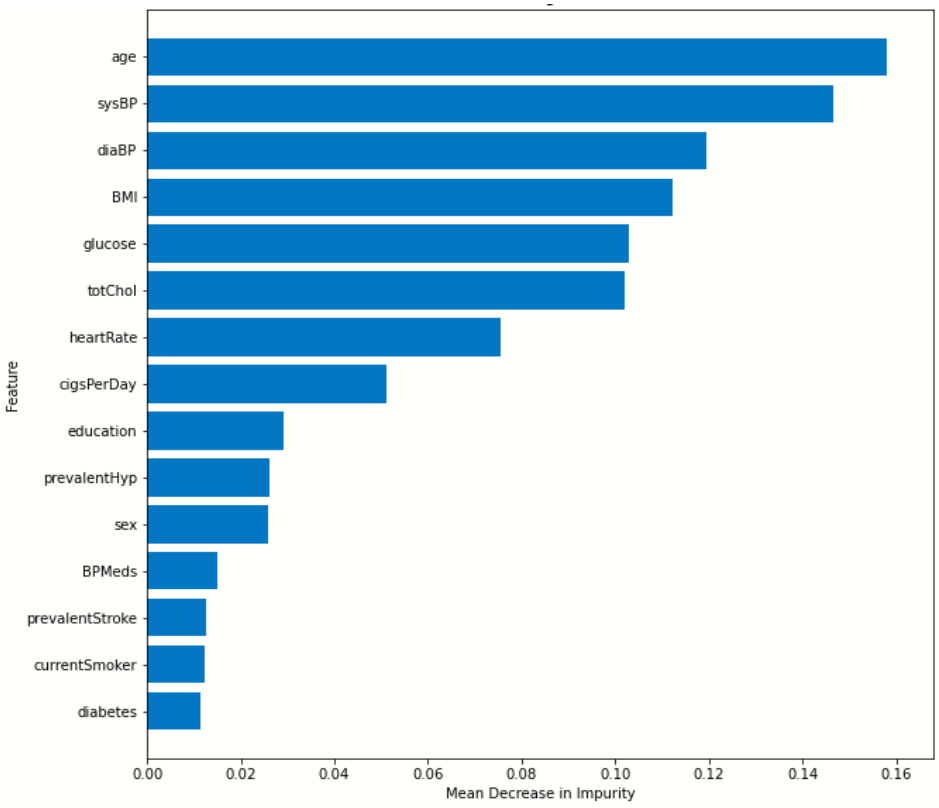


**Fig. 1.** Variable Ranking of RF (optimised hyperparameters) model for CHD Risk Prediction.

## 3.2 MetS Prediction

### *3.2.1 Characteristics of the study population*

The dataset used in the latter part of the study presented by this paper is from the NHANES and is publicly available on the Data World website. A total number of 6 781 participants

had enough data for inclusion in our study. The mean (std) age of participants was 39.10 (21.82) years, while 3 412 (50.31%) were male and 3 369 (49.69%) were female. 1061 participants (15.65%) had MetS, with the mean (std) age being 50.24 (19.97) where 531 (50.05%) are male and 530 (49.95%) are female.

### *3.2.2 Comparison of prediction models*

**Table 4.** Performance of MetS prediction models.

| Model | AUC-ROC | AUC Change |
|---|---|---|
| Random Forest (ignoring hyperparameters) | 0.753 | Baseline model |
| Random Forest (optimised hyperparameters) | 0.858 | +10.5% |
| AutoPrognosis (single pipeline) – using all variables | 0.773 | +2.0% |
| AutoPrognosis (ensemble pipeline) – using all variables | 0.851 | +9.8% |

The AUC-ROC values for the different models under evaluation for MetS prediction are shown in Table 4. The Random Forest model where hyperparameters were ignored during training (AUC-ROC: 0.753) was used as the baseline model for performance evaluation. For the other RF model (AUC-ROC: 0.858), the hyperparameters were optimized using the out-of-bag error value. This model significantly outperformed the baseline model. AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.851) significantly outperformed AutoPrognosis (single pipeline, using all core variables) (AUC-ROC: 0.773). Both AutoPrognosis models outperformed the baseline model. The RF model with optimized hyperparameters had the best overall performance, but only by a small difference (0.007) to the AutoPrognosis (ensemble pipeline) - using all variables.

### *3.2.3 Variable Importance*

Figure 2 displays a list of all variables ranked based on their contribution to the RF model with optimized hyperparameters. The importance scores are based on their mean decrease in impurity.

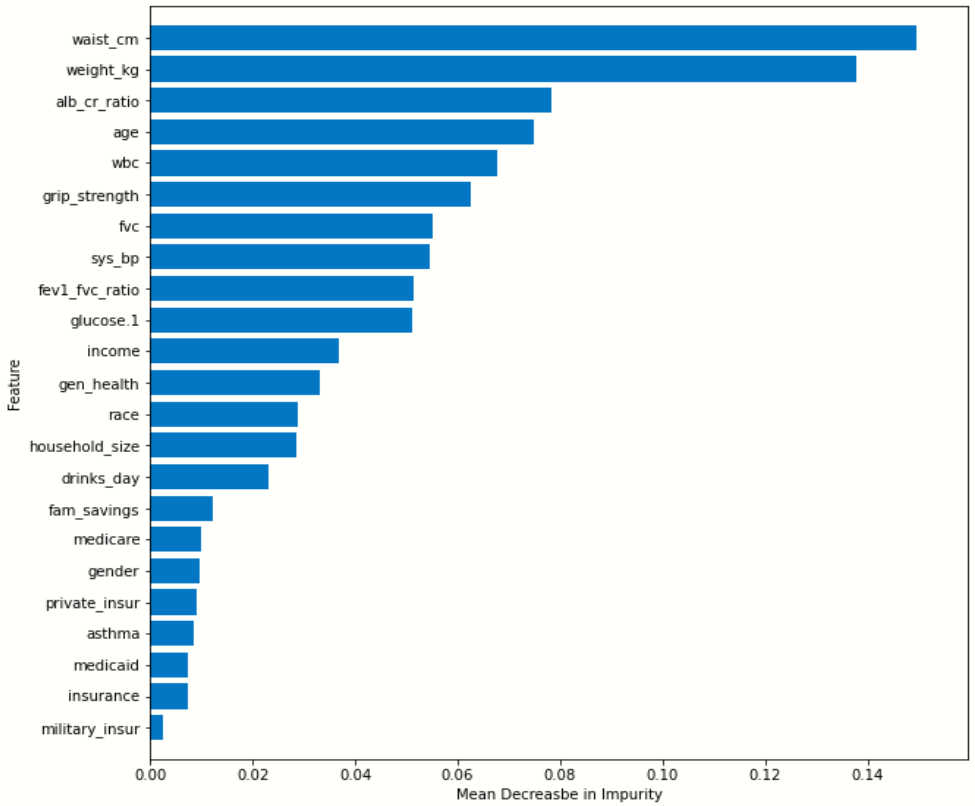


**Fig. 2.** Variable Ranking of RF (optimised hyperparameters) model for MetS Risk Prediction.

## 3.3 Performance of RF against AutoPrognosis on different sample sizes of data

Figure 3 displays the performances, indicated by the AUC-ROC value, of four different models which were each built using the NHANES dataset used in the previous section for predicting the presence of MetS. Once again, the four models are based on AutoPrognosis with a single pipeline, AutoPrognosis with an ensemble pipeline, Random Forest with default hyperparameter values and Random Forest with optimized hyperparameters values, respectively. All of these models had satisfactory performance results, on all the different sample sizes of data as shown in figure 1. The Random Forest model with optimized hyperparameters values had the best overall performance, followed by AutoPrognosis with an ensemble pipeline, then AutoPrognosis with a single pipeline and finally the Random Forest model with default hyperparameter values.

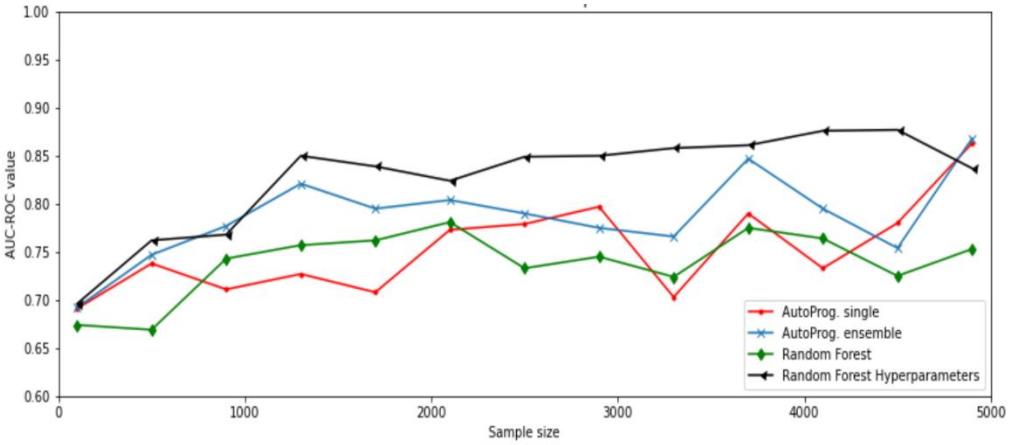


**Fig. 3.** AUC-ROC value vs Sample Size.

## 4 Discussion

CVD accounts for approximately 32% of all deaths worldwide, making it the leading cause of mortality globally, with CHD being the most common type of CVD [16]. CHD often results in the development of chest pains or arrhythmia and can even progress to the occurrence of a heart attack or heart failure. However, clinicians reiterate that 80% of heart attacks and CVD related events can be prevented [16]. MetS is a collection of conditions that can increase a person's risk of developing a CVD and/or diabetes. The individual components of MetS are recognized as CHD risk factors and are often included as some of the core variables used in models developed for CHD risk prediction. Predicting future CHD risk and MetS occurrence at an individual level, population level and in specific subgroups of the population will provide useful information for policymakers and healthcare authorities about these risks. CHD risk prediction and MetS detection can also motivate individuals to adjust their lifestyle choices, behaviours, and habits.

In this study, several ML models were built based on RF and AutoPrognosis, for both 10-year CHD risk prediction and MetS prediction. The CHD risk prediction model was built using a dataset of 3 658 records and 16 variables from the FHS, while the MetS prediction model was built using a dataset of 6 781 records and 24 variables from the NHANES. Our study discloses a number of significant points. Firstly, for CHD risk prediction both types of ML models significantly outperformed the FRS model based on conventional CVD risk factors. Therefore, using ML methods for clinical applications will be beneficial. Secondly, for CHD risk prediction, the RF model with optimized hyperparameters (AUC-ROC: 0.728) outperformed the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.714) but only by a very small margin. Thirdly, the same was true for MetS prediction, the RF model with optimized hyperparameters (AUC-ROC: 0.858) outperformed the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.851) but only by a very small margin. Among the conventional risk factors for MetS, at the top of our variable rankings are: *'alb_cr_ratio'* (ratio of albumin to creatine in urine; this measure is used to identify kidney disease that occurs due to complications of diabetes), *'wbc'* (white blood cell count), *'fvc'* (forced vital capacity which the maximum amount of air that can be exhaled forcibly after inhaling fully) and *'fev1_fvc_ratio'* (ratio of the amount of air expelled on one second to forced vitality capacity).

Expeditious advances in ML and the recent abundance of digitized healthcare data have resulted in a growing number of applications for ML within healthcare. Medical repositories often have vast quantities of data which has not been mined effectively. ML methods, such as those based on the RF and AutoPrognosis, are effective when working with large volumes of data to efficiently establish variable importance, the relationships among the variables, and make predictions.

Multiple comparative studies have been conducted to evaluate the difference in the performances between standard ML models and traditionally used CVD risk prediction models. One such study was performed by [10] on the Chinese population (30 variables; 29930 records), where an RF model was compared to models based on Naïve Bayes, Ada Boost, Classification and Regression Tree (CART), Bagged Trees and a multivariate regression model. Results showed that Random Forest was superior to the other models with an AUC of 0.787. Other popular studies include those found in [6, 8, 11] to name a few.

Furthermore, multiple other studies have been conducted with the aim of applying different ML methods to discover the optimal model for MetS prediction. One such study was performed by [15] on a Korean population (20 variables; 1 991 records), where the aim was to evaluate and compare the performances of nine different ML models based on Random Forest, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree, Logistic Regression, K-nearest neighbour, Multi-layer Perceptron, eXtreme Gradient Boosting (XGBoost) and 1D Convolutional Neural Network, respectively. Results showed that the XGBoost and RF models produced the best performances, with AUC-ROC values of 0.851 and 0.844, respectively. Other popular studies include those found in [12, 13, 14] to name a few.

To the best of our knowledge, our study is unique in that it is the first to conduct a comparative analysis for a performance evaluation between models based on RF against AutoPrognosis for MetS prediction.

Despite the progressive number of applications for ML models and techniques within prognostic research, often there exists a rift between the potential and actual practicality of these ML approaches. This is because clinicians without adequate knowledge and experience in data science are challenged to manually design and tune ML modelling pipelines before they can put them into use. AutoPrognosis was developed to circumvent this challenge, as it is an automated ML framework specifically designed for clinical prognosis [1]. Therefore, even though the RF model with optimised hyperparameters had the best overall performance in our study, AutoPrognosis may be the preferred method applicable for CHD risk prediction and MetS prediction, especially because there was a very modest difference in performance between the respective models.

In [2] a dataset from the UK Biobank for a cohort of 423 604 participants, with 473 available variables, was used to test whether AutoPrognosis is a more suitable option for CVD risk prediction in contrast to traditional approaches. An ML model developed using AutoPrognosis was compared to the Framingham Risk Score, a Cox Proportional Hazards (PH) model based on several well-established CVD risk factors, a Cox Proportional Hazards (PH) model based on all 473 variables from their dataset, and some standard ML models. It was depicted that AutoPrognosis improves CVD risk prediction in the population from the UK Biobank.

Findings from [8] suggest that all of the ML algorithms (Random Forest, Logistic Regression, Gradient Boosting and Neural Networks) they tested outperformed the well-established ACC/AHA ten-year risk prediction model. However, unlike most studies, they found that their highest performing ML model was based on Neural Networks as opposed to Random Forests. The size and dimensionality of the dataset used to train a model will often affect the performance of the algorithm the model is based on. Therefore, the choice of the best ML model and tuning of its hyperparameters are important for ensuring the possible advantages of ML applications. AutoPrognosis automates these processes, making it more easily useful for ML applications in the clinical space.

However, in [1] the performance of AutoPrognosis was evaluated using datasets with at least 30 000 patient records. Sometimes obtaining data that large and training models on it can be challenging and computationally expensive. Therefore, evaluating the performance of AutoPrognosis on significantly smaller datasets is essential as some ML models, such as Neural Networks, are data-hungry and require large quantities of data during training in order to produce satisfactory performance results. To the best of our knowledge, our study is unique in that it is the first to conduct a comparative analysis for a performance evaluation between models based on RF against AutoPrognosis on different sample sizes of data. Our results concluded that AutoPrognosis with a single pipeline, AutoPrognosis with an ensemble pipeline, Random Forest with default hyperparameter values and Random Forest with optimized hyperparameters values all had satisfactory performance results, on all the different sample sizes of data. The Random Forest model with optimized hyperparameters values had the best overall performance, followed by AutoPrognosis with an ensemble pipeline, then AutoPrognosis with a single pipeline and finally the Random Forest model with default hyperparameter values.

## 5 Limitations

The performance of our models was evaluated on relatively small datasets, as both datasets had less than 10 000 records. A larger dataset may produce better AUC results.
The performance of AutoPrognosis on different sample sizes of data, ranging from 100 to 4900, was tested. However, once again these are relatively small datasets. Further experimentation can be done on larger datasets of different sample sizes.
The two datasets used in our study were both imbalanced. Applying resampling techniques to deal with the class imbalance may lead to improved model performances. This can be applied in future work related to this study.

## 6 Conclusion

In this study, we built two types of 10-year CHD risk prediction models using data from the FHS. These two models were based on ensemble machine learning methods, namely Random Forests and AutoPrognosis. As a baseline model for performance evaluation, we used the FRS model. Results showed that the RF model with optimized hyperparameters had the best overall performance. As a novel study, data from the NHANES was used to assess the predictive performance of RF against AutoPrognosis for determining the presence of MetS. The RF model with optimized hyperparameters had the best overall performance.
The performance of RF a AutoPrognosis on different sample sizes of data, ranging from 100 to 4900, was also tested. The Random Forest model with optimized hyperparameters values had the best overall performance, followed by AutoPrognosis with an ensemble pipeline, then

AutoPrognosis with a single pipeline and finally the Random Forest model with default hyperparameter values.

## References

1. A. Alaa, M. Schaar, PMLR **80**, 139-148 (2018)
2. A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H.F. Rudd, and M. van der Schaar, PloS ONE **14 (5),** e0213653 (2019)
3. L. Breiman, Machine Learning **45 (1)**, 5-32 (2001)
4. R.B. D'Agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, Circulation **117 (6),** 743-753 (2008)
5. D.C. Goff, D.M. Lloyd-Jones, G. Bennett, S. Coady, R.B. D'Agostino, R. Gibbons, P. Greenland, D.T. Lackland, D. Levy, C.J. O'Donnell, J.G. Robinson, J.S. Schwarts, S.T. Shero, S.C. Smith, P. Sorlie, N.J. Stone, P.W.F. Wilson, JACC **63 (25 Part B),** 2935-2959 (2014)
6. B.A. Goldstein, A.M. Navar, R.E. Carter, Eur. Heart J. **38 (23)**, 1805-1814 (2017)
7. D.M. Lloyd-Jones, Circulation, **121 (15)** 1768-1777 (2010)
8. S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, PloS ONE **12 (4)**, e0174944 (2017)
9. P.W.F. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, W.B. Kannel, Circulation **97 (18)**, 1837-1847 (1998)
10. L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, J. Yan, Scientific Reports **10 (1)**, 1-8 (2020)
11. S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, T. Zhu, ICBDA, 288-232, IEEE (2017)
12. G.O. Gutiérrez-Esparza, O.I, Vázquez, M Vallejo, J. Hernández-Torruco, Symmetry **12 (4)**, 581 (2020)
13. E.K. Choe, H. Rhee, S. lee, E. Shin, S. Oh, J. Lee, S.H. Choi, Genomics Inform. **16 (4)**, e31 (2018)
14. C. Yu, Y. Lin, C. Lin, S Wang, S Lin, S.H. Lin, J.L Wu, S Chang, JMIR medical informatics **8 (3)**, e17110 (2020)
15. J. Kim, S. Mun, S. Lee, K. Jeong, Y. Baek, BMC Public Health **22 (1),** 644 (2022)
16. World Health Organisation, *Cardiovascular diseases (CVDs)*, who, June 11, 2021. https://www.who.int/news-room/ fact-sheets/detail/cardiovascular-diseases-(cvds)
17. Dileep, Logistic Regression To predict heart disease, kaggle, Accessed May 11, 2021, https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression
18. *Open Data Project with NHANES 2011-2012 Data*, data.world, Accessed February 05, 2022, https://data.world/rhoyt/librehealth-educational-ehr/workspace/file?filename=Merged_Unique_Names_V2.csv
19. Centers for Disease Control and Prevention, *National Health and Nutrition Examination Survey*, cdc, Accessed May 01, 2022, https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
20. National Heart, Lung, and Blood Institute, *Framingham Heart Study (FHS)*, nhlbi, Accessed May 03, 2022, https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs

21. R.B. D'AgostinoSr, R.S. Vasan, Michael.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, Circulation, **117 (6)**, 743-753 (2008)
22. A.M. Alaa, *AutoPrognosis: Automated Clinical Prognostic Modelling via Bayesian Optimization*, github, December 21, 2019, https://github.com/ahmedmalaa/AutoPrognosis
23. *Open Data Project with NHANES 2011-2012 Data*, data.world, Accessed February 05, 2022, https://data.world/rhoyt/librehealth-educational- https://data.world/rhoyt/librehealth-educational-ehr/workspace/file?filename=Codebook_NHANES_2011_2012.xlsx