Machine Learning and Statistical Approaches for Classification of Risk of Coronary Artery Disease using Plasma Cytokines.

Seema Singh Saharan[1], Pankaj Nagar[2], Kate Townsend Creasy[3], Eveline O. Stock[4], James Feng[5], Mary J. Malloy[6], John P. Kane [7].

Abstract

Background

As per the 2017 WHO fact sheet, Coronary Artery Disease (CAD) is the primary cause of death in the world, and accounts for 31% of total fatalities. The unprecedented 17.6 million deaths caused by CAD in 2016 underscores the urgent need to facilitate proactive and accelerated pre-emptive diagnosis. The innovative and emerging Machine Learning (ML) techniques can be leveraged to facilitate early detection of CAD which is a crucial factor in saving lives. The standard techniques like angiography, that provide reliable evidence are invasive and typically very expensive and risky. In contrast, ML model generated diagnosis is non-invasive, fast, accurate and affordable. Therefore, it can be used as a supplement or precursor to the conventional methods. This research demonstrates the implementation of K Nearest Neighbor (k-NN) and Random Forest ML algorithms to achieve a targeted "At Risk" CAD classification using an emerging set of 35 cytokine biomarkers that are strongly indicative predictive variables that can be potential targets for therapy. To ensure better generalizability, mechanisms such as data balancing, k-fold cross validation for hyperparameter tuning, feature selection via feature importance identification were integrated within the models.

Results

A total of 5 classifiers were developed, with two built using 35 cytokine predictive features and three built using a subset of cytokines, selected by variable importance techniques namely Random Forest, ReliefF and Boruta. The best Area under Receiver Operating Characteristic (AUROC) based accuracy of .99 was achieved by the

[1] M.Phil., Corresponding Author, Research Scholar, Department of Statistics, University of Rajasthan, Jaipur, Voluntary Data Scientist UCSF Kane Lab, San Francisco, Part Time Lecturer, UC Berkeley Extension.
Email: ssaharan9@gmail.com , seema.saharan9@berkeley.edu
[2] Ph.D., Associate Professor, Department of Statistics, University of Rajasthan, Jaipur.
Email: pnagar121@gmail.com
[3] Ph.D., Cardiovascular Research Institute, Department of Medicine, University of California, San Francisco.
Email: kate.creasy@ucsf.edu
[4] M.D., Cardiovascular Research Institute, Department of Medicine, University of California, San Francisco.
Email: eveline.stock@ucsf.edu
[5] B.S., Cardiovascular Research Institute, Department of Medicine, University of California, San Francisco.
Email: james.feng@ucsf.edu
[6] M.D., Cardiovascular Research Institute, Departments of Medicine and Pediatrics, University of California, San Francisco.
Email: mary.malloy@ucsf.edu
[7] M.D., Ph.D. Cardiovascular Research Institute, Department of Medicine, Department of Biochemistry and Biophysics, University of California, San Francisco.
Email: john.kane@ucsf.edu

Random Forest classifier with 35 cytokine biomarkers. The second-best AUROC accuracy was achieved by the k-NN model using cytokines selected by the Random Forest variable importance selection mechanism.

Conclusions

Presently, as large-scale efforts are gaining momentum to enable early, fast, reliable, affordable, and accessible detection of individuals at risk for CAD, the application of powerful ML algorithms can be leveraged as a supplement to the conventional treatments such as angiography. The early detection can be further improved by incorporating 65 novel and sensitive cytokines biomarkers. Investigation of the emerging role of cytokines in CAD can materially enhance the detection of risk and the discovery of mechanisms of disease that can lead to new therapeutic modalities.

Keywords

CAD, ML, k-NN, Random Forest, Distance Metrics, k-fold Cross Validation, Classification, AUROC, Plasma cytokines, ROSE, ReliefF, Boruta.

Background

Introduction

Cardiovascular disease is the leading cause of death in Europe and North America [1] [2] which underscores the need for incorporation of novel emerging risk factors to improve prediction of risk, enabling early diagnosis and personalized management. The power of ML algorithms like k-NN and Random Forest can be harnessed to extract patterns to inform health related decision making. This paper expatiates the exploratory juxtaposition of k-NN and Random Forest by varying a broad spectrum of tuning parameters and incorporating feature importance in conjunction with k-fold cross validation, a powerful resampling technique that overcomes the issue of overfitting, ensuring better generalizability of the model [3]. Cytokine feature importance was determined by numerous methods such as Random Forest, ReliefF and Boruta. Due to the limitations of data availability, before creating the model and final prediction the data was augmented and balanced by Random Over-Sampling Examples technique(ROSE).ROSE, available from the Comprehensive R Archive Network (https://cran.r-project.org/web/Packages/ROSE/index.html) simulates balanced synthetic data by smoothed bootstrap approach. We used 35 plasma cytokines as novel biomarkers to improve classification in patients with or without clinical coronary disease. This approach promises to identify mechanisms of disease with cytokine targets not previously recognized, and to improve early detection of individuals at risk. Cytokines are proteins generated by the immune system in response to cell signals. They act as messengers for other cells by targeted activation of receptors and trigger downstream, signaling. Common cytokines include lymphokines, chemokines, interferons, interleukins

etc. that respond to environmental signals triggering pro or anti-inflammatory cascades [4] [5] [6]. Cytokines are known to be involved in the development and progression of CAD [7].

Review of Related Work

There are very few prior studies that have used ML algorithms to differentiate CAD versus Control by using cytokines [8] [5] as predictive features emphasizing the importance of the current study. These studies also underscore the importance of ML techniques and how they can be harnessed to predict CAD versus Control as a supplement to traditional methods to direct treatment. k-NN is very popular within the landscape of ML algorithms due to its interpretability and distance metric options for measuring the similarity within predictor features resulting in the target feature classification. This algorithm has been used across a broad spectrum of domain areas including medical diagnosis of which some prominent research is discussed here. Random Forest [3], an ensemble of trees is also a very innovative technique that has been empirically proven to generate high predictive accuracy specially for a high dimensional data.

The ML techniques to predict CAD versus Control can be used in conjunction with traditional methods or even independently to facilitate early, fast, affordable, accessible, noninvasive prediction without compromising accuracy. In light of the expensive and risky nature of some of the traditional methods utilized to diagnose CAD, ML has gained impetus to offset the current limitations.

Alizadehsani, Roohallah, et al. [9] delineates the usage of ML algorithms in conjunction with angiography to obviate its disadvantages which include costs, complications, and after-effects of the invasive method. By the usage of ML algorithms, the researchers identified the subjects that were at a very high risk and then these were referred for angiography. The study used laboratory and echocardiographic data to diagnose the stenosis of each artery separately. Bagging and C4.5 classification algorithms were used to obtain an accuracy of 79.54%,61.46% and 68.96% for diagnosis of stenoses of left anterior descending (LAD), left circumflex (LCX) and right coronary artery (RCA), respectively.

Mastoi, Qurat-ul-ain, et al. [10] investigated the automation of CAD diagnosis by the ML algorithms such as k-NN and SVM. The motive of the automation was to find alternative ways to time consuming and expensive procedures that patients must undergo. The researchers developed less expensive and effective alternatives to medically prescribed tests such as angiography, nuclear scan, and C-reactive protein measurement that are expensive and require technical expertise. The features used for this classification were non-invasive biomarkers such as electrocardiography (ECG), photoplethysmography (PPG), and phonocardiography (PCG). Other studies used clinical parameters such as age, blood pressure, and smoking habit. The best prediction accuracy of 99% was achieved by SVM.

Hampe, Nils, et al. [11] have proposed the exploration of cardiac computed tomography (CT) visualization by ML algorithms. CT generates detailed high spatial resolution with hundreds of slices which are under-utilized due to paucity of trained cardiac imagers and overwhelming workload for medical professionals. ML algorithms can surmount the obstacles of manual diagnosis by achieving accurate and fast results which might lead to additional secondary diagnosis as well. The survey explored and documented ML algorithms augmenting CAD detection and characterization spanning the past ten years. The conclusion states that despite the challenges pertaining the implementation of ML within the clinical scope , the power of novel ML algorithms is providing an impetus to gain substantive insights in CAD classification.

Martin-Isla, Carlos, et al. [12] investigated the uses of ML algorithms for image based diagnosis of CAD, which have accomplished deeper qualification and superior diagnosis due to the generative nature of ML algorithm that learn from past predictions. Furthermore, the potential of ML algorithm for CAD detection is emphasized by the extensive literature related to the domain. The ubiquitous implementation of ML algorithm has gained momentum because of the availability of high computational power and the outstanding prediction accuracy which is an improvement of the current qualitative assessment of images and crude quantitative measures of cardiac structure and function. The ML algorithm can build a holistic framework encompassing not just images but also other informative features to obtain credible insights and early detection which will result in saving lives.

Yu, Linghua, et al. [6] studied hand, foot, and mouth disease (HFMD) prevalently found in the Asia-Pacific regions. The research participants implemented Random Forest to distinguish the HFMD disease group from the controls using 26 significant cytokines as predictor features. The findings of the research showcased correlation between enteroviral infection, genotype, and clinical presentation. The Random Forest algorithm achieved a final AUROC value of .91, demonstrating its excellent partition efficacy. This shows that cytokines are sensitive and powerful predictive biomarker features.

Struck, et al. [13] employed cytokine predictors, using Random Forest to differentiate malaria from a blood stream bacterial infection. The 7-15 cytokines used for the task were selected using ML classification techniques. The researchers used cytokines to offset the deficiency of a rapid malaria test not being able to differentiate serious malaria infection from asymptomatic malaria. This study exhibited a high disease status prediction accuracy of 88% that could provide directions to develop new point-of-care tests in Sub-Saharan Africa.

Kandhasamy et al [14] tackles the early prediction of diabetes, a commonly occurring disease in the contemporary context using important ML classifiers like Decision trees, k-NN, Random Forest and Support Vector Machines. Additionally, improved efficacy of non-noisy versus noisy data is also demonstrated by using performance metrics like Prediction Accuracy, Sensitivity and Specificity. To generate a more consistent dataset, missing values were replaced by the median value of the attribute across all observations.

Jabbar, Akhil, et al. [15] used an innovative approach for diagnosing heart disease, by combining k-NN algorithm with a genetic algorithm. The research empirically proved that the inclusion of a genetic algorithm within the folds of k-NN enhances the prediction accuracy thereby providing an inventive diagnostic approach.

Enriko, Ketut & Suryanegara et al [16] experimented and compared ML algorithms such as Naïve Bayes, Decision Tree and k-NN by only applying 8 biomarker features instead of the recommended 13 features. The 8 features were chosen because they were simple to measure and provided a better prediction accuracy for k-NN compared to Naïve Bayes and Decision Trees. The use of multiple algorithms provided a comparative analysis with regards to an evaluation measurement like prediction accuracy. k-NN and SVM achieved a prediction accuracy of 73% whereas Random Forest achieved a prediction accuracy level of 71%.

Faizal, Edi, and Hamdani. [17] extracted the early diagnosis of CAD based on feature similarity with predictive attributes such as age, gender, and symptoms. Specifically, the classification technique used for this research was the k-NN weighted via the Minkowski distance. To reduce the chance of error in terms of differentiation, it was decided that if the similarity index were less than .80, the diagnosis would be determined with the consultation of an expert. The performance metrics like Prediction Accuracy of 95.83% obtained by the application of this methodology were high.

Saini, Indu, et al. [18] studied the usage of k-NN for the detection of QRS complexes in ECG related data. The authors showed that prediction accuracy primarily depended on the value of k and the distance metric used for classification. Running experiments, proved that Euclidean distance and a value of k=3 in conjunction with 5-fold cross validation generated the best k-NN classifier. The prediction accuracy achieved was 99% which is remarkably high.

Ridker, Paul M, et al. [19] conducted CANTOS , a Clinical trial funded by Novartis (Canakinumab Anti-inflammatory Thrombosis Outcome Study) whose objective was to investigate the involvement of interleukin-1 β in the inflammation at a cellular level. The experiment entailed giving the men in the study a monoclonal antibody against interleukin-1 β .The anti-inflammatory therapy targeted the interleukin-1 β innate immunity pathway with canakinumab at a dose of 150 mg every three months. The study exhibited a significant lower rate of cardiovascular events than placebo independent of lipid lowering.

The current study endeavored to improve and extend the techniques in the published studies. The primary objective is to leverage the powerful algorithms like k-NN and Random Forest with the emerging cytokine biomarkers to obtain a better separability evaluated by the AUROC curve. This diagnosis mechanism is not a substitute for the conventional methods such as angiography but can be used to recommend more complicated tests for patients with more serious disease. The usage and comparison of multiple algorithms has proven an effective way to obtain a holistic view with regards to the classification. The current study incorporates this

comparison paradigm to present the [20] results. The empirical proof of enhanced performance by the use of hyperparameter tuning and cross fold validation [21] for Random Forest and k-NN directed the in-depth exploration conducted in the current research. Fine tuning hyperparameters in general has proven to be an effective optimization technique. The novel cytokine biomarkers that are indicators of inflammation [19] [22] [20] can be leveraged to make the identification of CAD risk more substantive and could be comprehensive targets for future therapies.

Methods

The data set is composed of biomarker levels for 104 individuals. Thirty-five cytokine biomarkers were measured for all in addition to the final target feature attribute which assigns individual to the CAD (39 individuals) or the Control (65 individuals) group. The feature space in the model incorporates 35 cytokine biomarkers that helped quantify the similarity and finally the classification of CAD or Control.

This study was approved by the UCSF Institutional Review Board Committee on Human Research and conducted in accordance with the principles of the Declaration of Helsinki. All subjects provided written informed consent prior to participation. For this study, blood samples were collected from male (43.3%) and female subjects, ages 18 to 65 (median age = 42) with diagnosed CAD, and from age, and sex-matched controls. CAD subjects had a previous history of myocardial infarction, angiographically diagnosed CAD, or previous coronary artery bypass graft surgery. Control subjects had no history or clinical evidence of CAD. Exclusion criteria included current or prior treatment for autoimmune disease and/or cancer, diabetes, tobacco use, NSAID use prior to blood collection, post-menopausal women, and age over 65. None of the subjects were receiving lipid lowering medications. Blood was drawn into EDTA collection tubes and immediately stored on ice. Samples were centrifuged to separate plasma which was aliquoted and stored at -80° C until use. Samples were thawed on ice and assayed for cytokine content with a human 35-plex ELISA assay (ThermoFisher/Life Sciences) according to the manufacturer protocol. The raw data were analyzed by xPonent software and were expressed in pg/mL using the standard curve for each cytokine. Table 1 describes the clinical demographic profile of the subjects in the study.

Table 1: Clinical Demographic Profile. A retrospective collection of plasma samples from patients with diagnosed coronary artery disease (CAD) and healthy controls.

Gender	CAD (n=39)	Control (n=65)	Total (n=104)
Male	19	26	45(43.27%)
Female	20	39	59(56.73%)
Total	39(37.5%)	65(62.5%)	104

Pre-processing Steps

Prior to running the classifier on the testing set, the attributes/features of the data were normalized to prevent the predictive features with larger values from dominating the features having smaller values which would result in a

biased classification. Additionally, the data were checked and remediated for discrepancies such as null values or outlier values by consulting with domain experts. Due to the availability of a small data set which was significantly imbalanced, the data were incremented synthetically by smoothened bootstrap mechanism which prevented overfitting in training phase and translated in better generalizability during the testing phase. By using the ROSE Package from the R programming language, the data were synthetically increased to a size of 1000 and balanced with CAD accounting for 52% of the cases and Controls 48% .The ROSE Package helps achieve this by simulating smoothened bootstrap approach. The data was scaled by z-score standardization.

The final balanced data composed of 52% CAD and 48% Control is reasonable for implementing k-NN and Random Forest .The augmentation of the data to 1000 observations in conjunction with data balancing allows for a conventional 75% -25% split toward the training and testing partitions. This segmentation will allow for enough observations to be included within the training set to avoid underfitting the model. Additionally, 10-fold cross validation was implemented for choosing hyperparameters to prevent biased predictions. The predictive feature space consisted of 35 cytokines therefore for some of the classifiers feature selection was conducted to avoid underfitting or overfitting the model. The optimal balance of bias variance (underfitting versus overfitting) tradeoff was achieved by these strategies.

k-NN K-Nearest Neighbor

k-NN, a supervised ML algorithm was initially proposed by Fix and Hodges [23]. It is based upon the similarity paradigm, indicating that the classification of unlabeled examples is differentiated by means of distance metrics and are finally ascribed the class of k($k \geq 1$) nearest neighbors. k-NN, by virtue of being non-parametric in nature does not make assumptions regarding the underlying data distribution, therefore making it less restrictive and a more powerful classifier as compared to other popular ML algorithms. k-NN is a versatile algorithm that can be used for classification as well as prediction via Regression.

k-NN is a lazy learner therefore does not create a learning model, but instead every new testing instance is iterated through the training data to decide upon its class label. An increase of data instances causes a higher computational complexity due to the lack of the abstraction phase. k-NN also has the disadvantage that it does not predict well for data that are noisy or have outliers. Despite the caveats, the availability of computational power in contemporary context as well as the hyper-parameters that can be tuned for k-NN, allow it to be leveraged to adequately classify testing examples in a reasonable amount of time.

In the past as well as in the present, a plethora of complex medical research has applied k-NN [16] to achieve optimal diagnostic prediction. k-NN is prevalently used for detecting genetic diseases, conducting facial recognition, and generating music recommendation . The choice of this algorithm stems from the fact that even

though classification can be slow, k-NN is fundamentally a simplistic algorithm which typically uses numeric predictor features, is easily comprehendible and outperforms many of the more complex ML algorithms.

1. Parameter K (Number of Neighbors) Fine Tuning

The optimal value of hyper parameter k is decided by empirically initiating the algorithm with k=1 and iteratively incrementing k until the classifier's error rate is minimized. This technique helps prevent under fitting as well as overfitting of the testing data thereby balancing the bias variance trade-off. If k is too small, there is a reasonable possibility that an outlier will affect the classification and if k is too large the similarity neighborhood might incorporate several deviant classes. For a noisy dataset, where the nearest neighbors vary widely in their distances, closest neighbors are more reliable for class label characterization and are given priority weighting by the process of majority vote.

2. Similarity Distance Metric Parameter

To compute the k-NN similarity index using the contextual feature space, the distance between two feature vectors is prevalently measured using Euclidean distance. Much of the current analysis was implemented using the R Package caret [24], (available from the Comprehensive R Archive Network at http://CRAN.R-project.org/Package=caret) and R Package ggplot2 [25] , (available from the Comprehensive R Archive Network at https://cran.r-project.org/web/Packages/ggplot2/index.html).

The Euclidean distance was implemented for the current research which is derived for a L_2-norm.

2.1 Minkowski distance

Mathematically, the distance $d(x, y)$ in a D-dimensional feature space between two points

$x = (x_1, x_2, x_3 \ldots \ldots x_D)^T$ and $y = (y_1, y_2, y_3 \ldots\ldots y_D)^T$ is represented as follows:

$$\mathbf{d(x, y) = ||x - y||_p = (\sum_i^D |x_i - y_i|^p)^{1/p}} \quad (1)$$

The L_p-norm is defined as the Minkowski distance where p is the factor depicting the norm.

2.2 Euclidean distance

If p=2, L_2-norm is defined as the Euclidean distance.

$$\mathbf{d(x, y) = ||x - y||_2 = (\sum_i^D |x_i - y_i|^2)^{1/2}} \quad (2)$$

Random Forest

The Random Forest classification model consists of many decision trees operating as an ensemble, that result in the target class with the majority vote assigned to the test example. The low correlation between models helps ensures that the composite classification of the ensemble outperforms any individual classification by offsetting the errors of each model. Bagging or Bootstrap Aggregation are used to implement diversity within the tree models Each model uses randomly sampled training data extracted with replacement that generates a distinct tree. This procedure does not allow replicating the training data since sub-setting a record cannot be chosen more than once.

Additionally, unlike a simple decision tree, the ensemble decision trees are forced to split on a node as per a randomly selected distinct feature, which might not be the best partition criterion resulting in low correlation amongst the differentiated parallel trees. The trees in the Random Forest are not only disparate regarding training data, but also regarding node split feature partition.

Optimization techniques

The optimizing mechanism of k-fold cross validation as well as inclusion of statistically significant cytokines were incorporated to enhance the final classification result.

1. k-fold Cross Validation

k-fold Cross Validation is a technique that optimizes the prediction ability of a model in the context of new unlabeled data consequently offsetting issues like overfitting or selection bias. The technique entails partitioning a dataset into k complementary subsets, implementing training of the model on k-1 subsets, and finally validating it on one partition. This study used k=10 to implement 10-fold cross validation.

2. Variable Importance by Random Forest

The variable selection accomplished by Random Forest, [26] by mathematically evaluating the mean decrease in node impurity. The importance of a variable $\boldsymbol{X_k}$ when prediction $\boldsymbol{Y}$ is evaluated by summing the weighted impurity reduction for all the $\boldsymbol{t}$ nodes. Gini function ($\boldsymbol{I}$) with an index $\boldsymbol{i}$ is mathematically represented as follows:

$$\boldsymbol{I}(\mathbf{X_k}) = \frac{\mathbf{1}}{\mathbf{M}} \sum_{\mathbf{m}} \sum_{\mathbf{t}} \frac{\mathbf{N_t}}{\mathbf{N}} \Delta \mathbf{i}(\mathbf{t}) \qquad \textbf{(3)}$$

Gini Index calculates the impurity of the classification. Gini Index varies between 0 and 1.

3. Variable Importance by ReliefF

ReliefF, an algorithm developed by Kira and Rendell [27] is a feature selection filter method that is notably sensitive to feature interactions. ReliefF identifies the feature scoring by finding differences between nearest neighbor instance pairs. This method also does not degrade due to noisy data.

If data consists of n instances and p features that belong to two classes then at each iteration of the algorithm, feature vectors closed to X are identified by the Euclidean distance. The closest same class instance is called "near hit" and the closest different class instance is called "near miss". Finally, the weight vector is calculated by the following formula

$$\mathbf{W_i} = \mathbf{W_i} - (\mathbf{x_i} - \mathbf{nearHit_i})^2 + (\mathbf{x_i} - \mathbf{nearMiss_i})^2 \qquad (\mathbf{4})$$

The weight of a feature decreases if it differs from that feature in the nearby instances of the same class more than nearby instances of the other class. Finally, the weight factor divided by m in m iterations is defined as relevance. Features are selected if their relevance factor is greater than a threshold. The threshold is defined by the formula $1/\sqrt{\alpha m}$ where α is the given confidence interval.

4. Variable Importance by Boruta

Boruta [28] is a wrapper around a Random Forest implementation which removes irrelevant features by a mechanism of identifying features that are less relevant then random probes determined by a statistical test. This was implemented by the Boruta R Package (available from the Comprehensive R Archive Network at http://CRAN.R-project.org/Package=Boruta). Random Forest generates classification by majority voting of multiple decision trees developed by bagging samples of training set. The loss of accuracy is computed by using random permutation of attribute values between instances for all the trees. The z-score is calculated and compared to a shadow attribute's z-score created for comparison by randomly shuffling values of the original attribute multiple times across different instances. The set of importance of the original attributes are ascertained by using shadow attributes as reference.

The importance of a shadow attribute can be nonzero only due to random fluctuations. Thus, the set of importance of shadow attributes is used as a reference for deciding which original attributes are truly important.

Classifier Experimental Framework

Across all classifier 10-fold cross validation ,data scaling and balancing was implemented by using the R caret Package. The first classifier experiment entailed applying the k-NN algorithm involving 35 cytokine predictor features with the Euclidean distance. The second classifier experiment involved the usage of k-NN algorithm, incorporating a subset of cytokines selected by Random Forest variable importance mechanism with the Euclidean distance. The third classifier was like the second except that the ReliefF method was used to ascertain variable importance of cytokines used for prediction. The fourth classifier had the same configuration as the second and the third except that it used Boruta method for variable importance. The fifth classifier implemented the Random Forest with 35 cytokines.

The graphical representation of the Classifier Experimental framework is provided in the following figure

(Fig. 1):

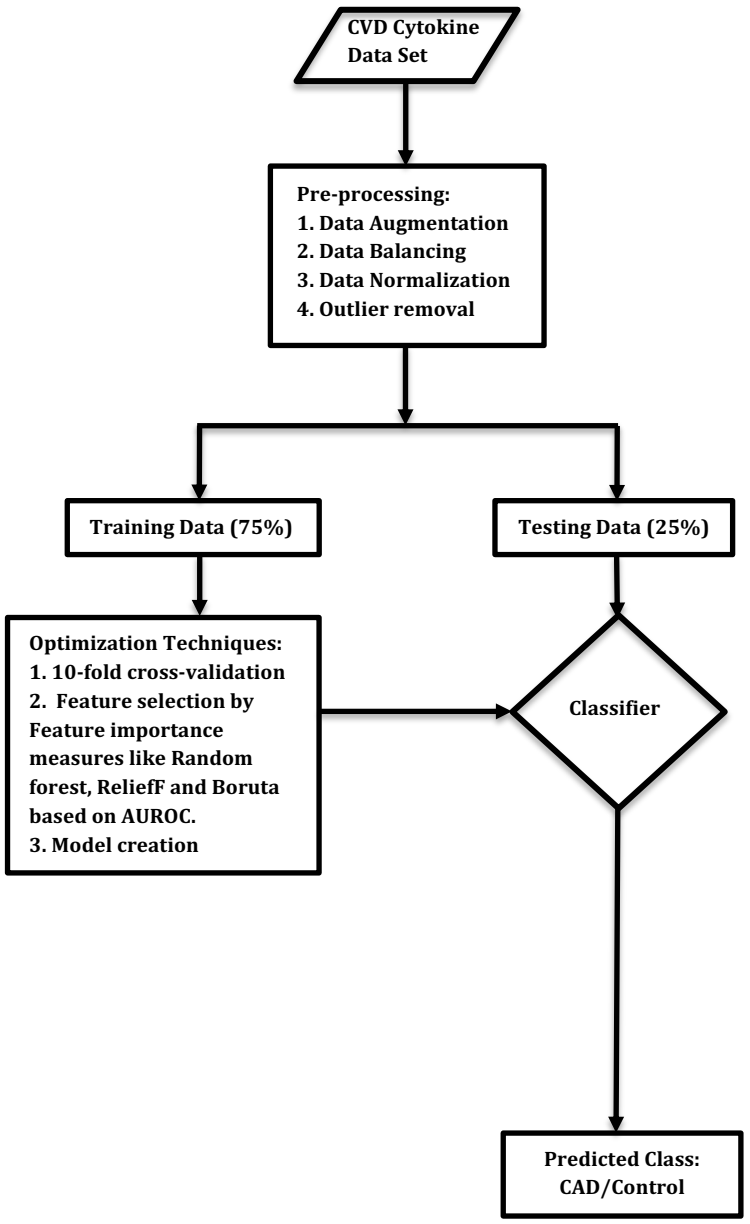

Fig. 1: The classifier framework implemented to identify "At Risk" Coronary Artery Disease (CAD) classification

Evaluation Measures

A versatile set of performance evaluation measures are available to obtain insights related to the efficacy of algorithms in terms of accuracy. For this study we used AUROC to obtain the accuracy of differentiating the CAD versus Control groups.

AUROC (Area under Receiver Operating Characteristic)

This is a standard measure that helps determine the degree of separability achieved by the relevant Classification algorithm. Higher AUROC showcases the algorithm's capability of accurately differentiating the instances into the target classes. The AUROC curve is created by plotting False Positive Rate (FPR) on the x-axis against the True Positive Rate (TPR) on the y-axis.

Results

Testing Results

The testing results obtained by running the algorithm on the test data are displayed in the following tables (**Tables 2-6**) and graphs (**Fig. 2-8**). The testing results applied the hyperparameter tuning metrics optimized by the resampling 10-fold cross validation results.

1. Classifier 1 Experiment

This classifier used k-NN with Euclidean distance measure and k=9 to classify the "At Risk" instances. For this classification 35 cytokines were used. The AUROC value of .95, representing the extent of separability of CAD versus Control was remarkable. The details related to AUROC and numeric measures for Classifier 1 are provided in the following table and graph. **(Table 2, Fig. 2).**

Table 2: Classifier 1 Experiment Results for the k-NN algorithm with 35 cytokines and k=9

Algorithm	Classification Criterion	Predictor Feature Space	AUROC	Prediction Accuracy	Sensitivity	Specificity
k-NN	Distance Measure: Euclidean with k=9	35 Cytokines	**0.95**	0.832	0.992	0.658

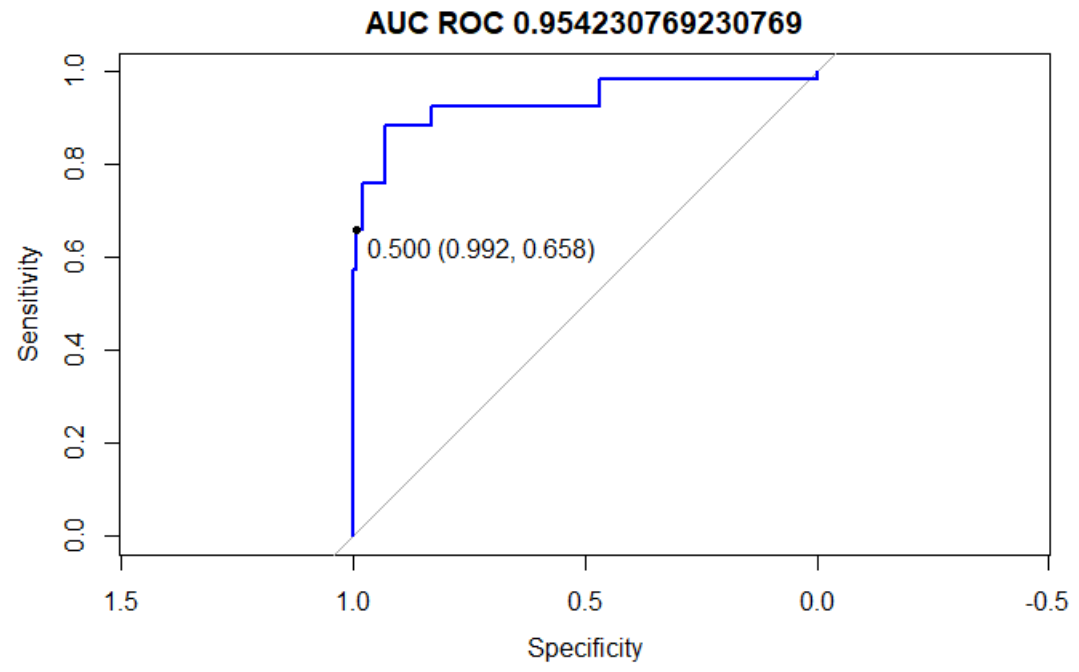

Fig. 2: Classifier 1 Experiment: AUROC curve for k-NN (Euclidean Distance) with a total set of 35 cytokines.

2. Classifier 2 Experiment

313 This classifier used k-NN with Euclidean distance for 10 cytokines, selected using the rf(Random Forest) variable
314 importance mechanism and k=9 .The AUROC value for this classifier was .96, which was slightly higher than
315 Classifier 1. The enhancement of AUROC accuracy can be attributed to the elimination of noisy data by the virtue
316 of dropping the 25 predictive features that were not highly indicative towards the classification. The information
317 regarding numeric results, AUROC and variable importance scores generated before the training phase are
318 provided in the following table and graph **(Table 3, Fig. 3, Fig. 4)**

Table 3: Classifier 2 Experiment Results for the k-NN algorithm with 10 cytokines and k=9

Algorithm	Classification Criterion	Predictor Feature Space	AUROC	Prediction Accuracy	Sensitivity	Specificity
k-NN	Distance Measure: Euclidean with k=9	10 Cytokines	**0.96**	0.828	0.9923	0.65

319

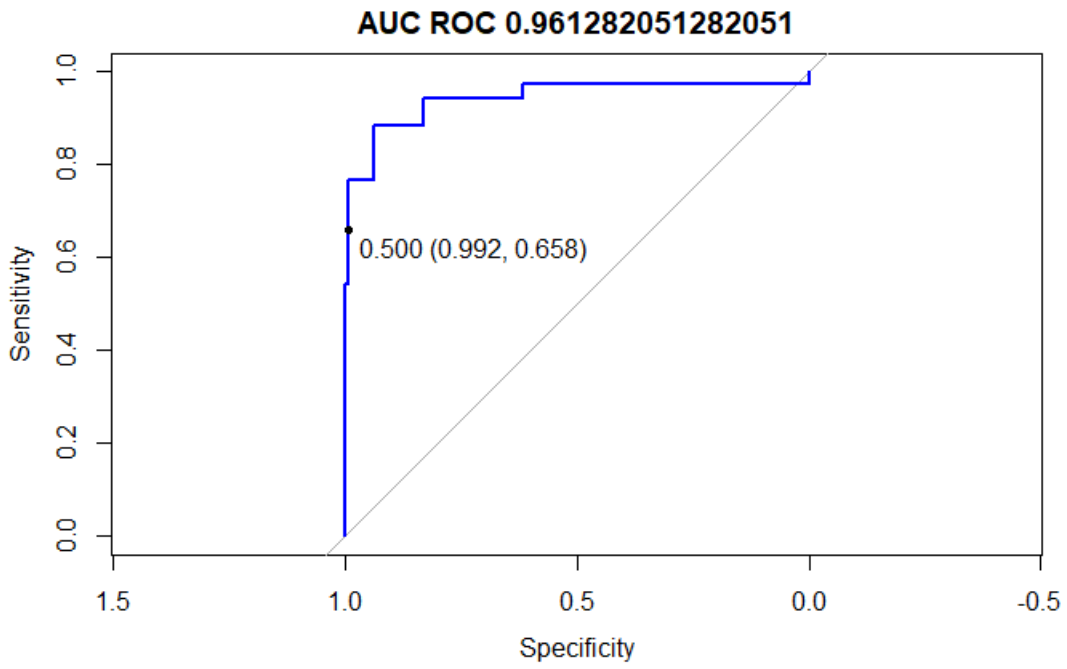

Fig. 3: Classifier 2 Experiment: AUROC curve for k-NN (Euclidean Distance) with 10 important cytokines identified by Random Forest mechanism.

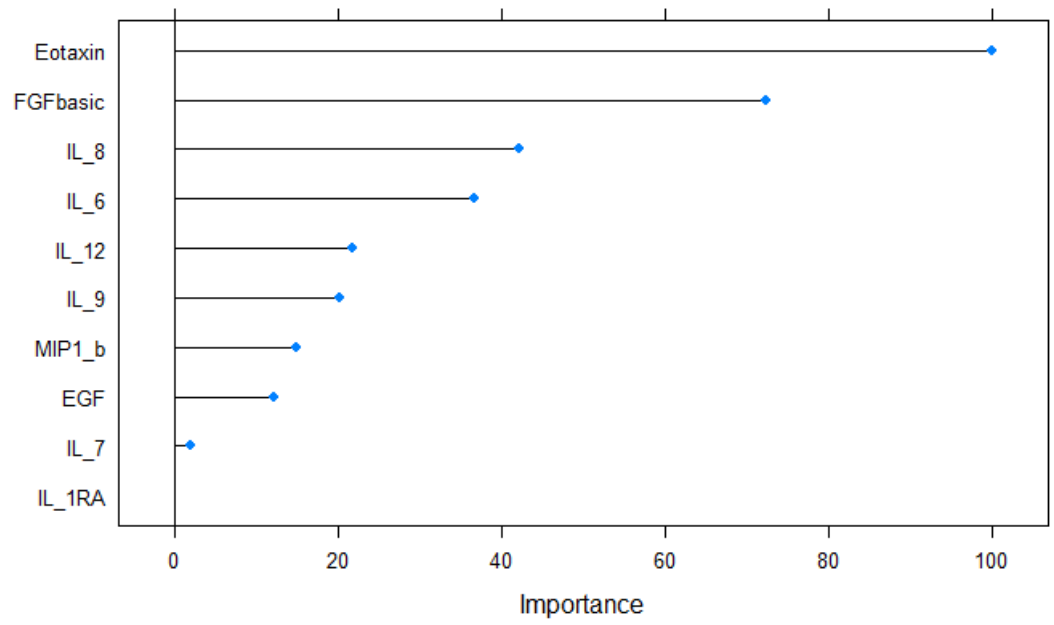

Fig. 4: Classifier 2 Experiment: Variable Importance scores of 10 best features determined by Random Forest algorithm before training the model. The top ten most important cytokines included Exotaxin , FGFbasic , IL_8 , IL_6 , 1L_12 , 1L_9 , MP1_b , EGF , IL_7 , 1l_1RA.

3. Classifier 3 Experiment

The third classifier applied k-NN and the importance of variables was ascertained by the ReliefF method. The ReliefF method identified the most important cytokines as follows: "IL_10", "FGFbasic", "HGF", "GCSF" "IL_1RA", "IL_7", "IL_2", "IL_9", "IL_8", "IL_5".

The experiment incorporated 10 cytokines with k=9 within the feature space. The AUROC obtained was equal to .935,which was slightly lower than the ones obtained for Classifier 1 and 2.The numerical details and AUROC are provided in the following table and graph **(Table 4, Fig. 5)**

Table 4: Classifier 3 Experiment Results for the k-NN algorithm with 35 cytokines and k=9

Algorithm	Classification Criterion	Predictor Feature Space	AUROC	Prediction Accuracy	Sensitivity	Specificity
k-NN	Distance Measure: Euclidean with k=9	10 Cytokines	**.93**	.91	.977	.767

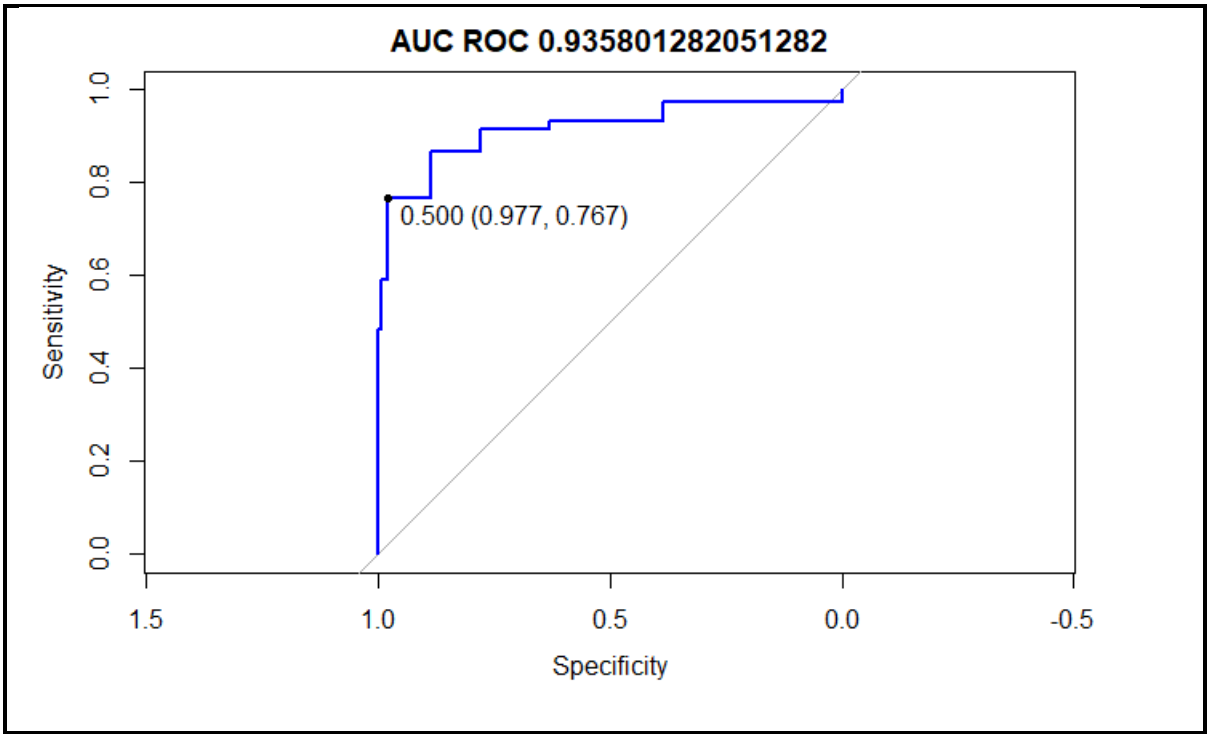

Fig. 5: Classifier 3 Experiment: AUROC curve for k-NN (Euclidean Distance) with 10 important cytokines identified by ReliefF. The 10 important cytokines identifed by ReliefF included : IL_10, FGFbasic, HGF, GCSF, IL_1RA, IL_7, IL_2, IL_9, IL_8, IL_5 .

4. Classifier 4 Experiment

In contrast to Classifier 3, this classifier used only important cytokines as predictor features identified by Boruta. The AUC of .92 was achieved, which was less than the AUROC for Classifier 1, Classifier 2, and Classifier 3. The details pertaining numerical measures and AUC for Classifier 4 are stated in the following table and graph **(Table 5, Fig. 6, Fig. 7)**

Table 5: Classifier 1 Experiment Results for the k-NN algorithm with 35 cytokines and k=9

Algorithm	Classification Criterion	Predictor Feature Space	AUROC	Prediction Accuracy	Sensitivity	Specificity
k-NN	Distance Measure: Euclidean with k=9	9 Cytokines	**.92**	.90	.992	.650

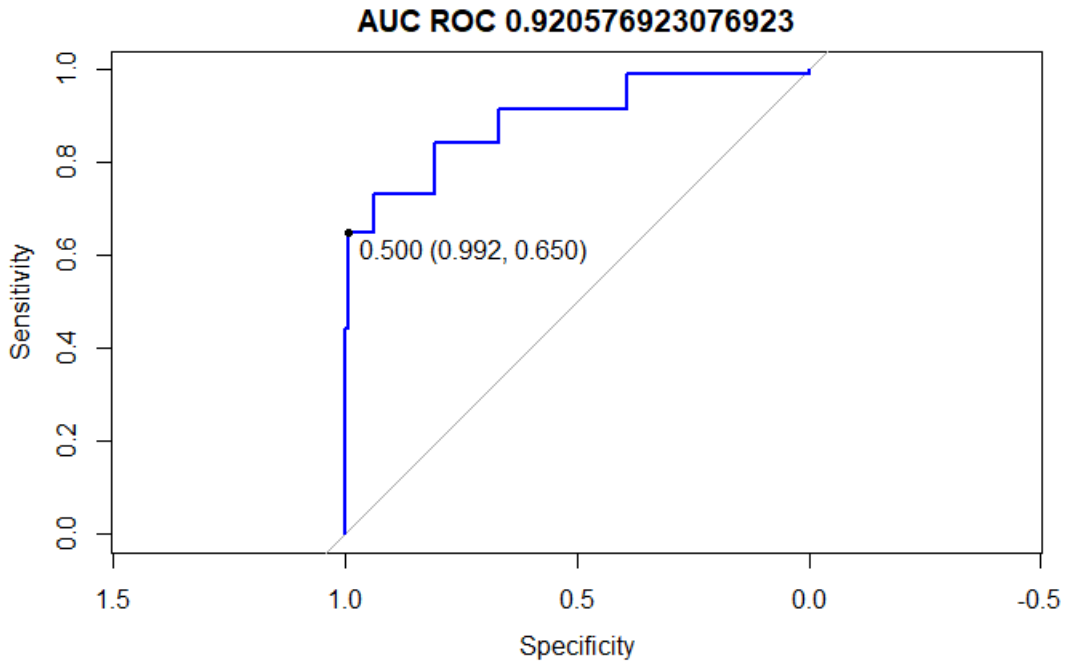

Fig. 6: Classifier 4 Experiment: AUROC curve for k-NN (Euclidean Distance) with 10 important cytokines identified by Boruta.

334

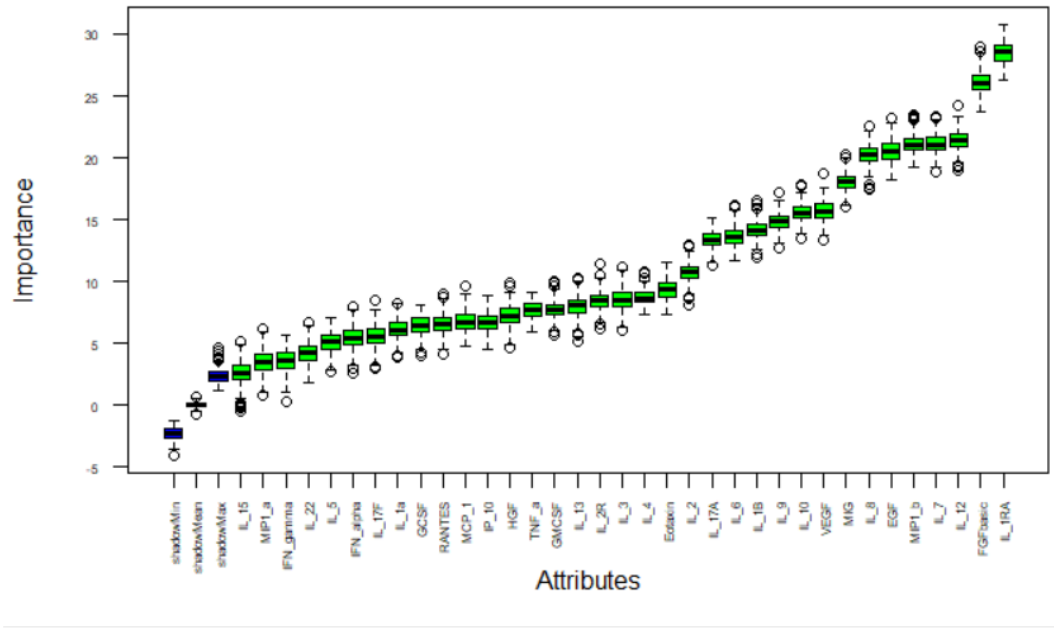

Fig. 7: Classifier 4 Experiment: Variable Importance scores of 10 best features determined by Boruta algorithm before training the model. The 10 important cytokines in ascending order are as follows: IL_1RA, FGFBasic, IL_12, IL_7, MP1_b, EGF, IL_8, MIG, VEGF, IL_10.

335 **5. Classifier 5 Experiment**

336 The Classifier 5 implemented Random Forest using 35 cytokines as predictor features. The AUROC of .99 was
337 extremely high and surpassed the accuracy achieved by the previous four classifiers. This accomplishment can be
338 accounted for by the underpinnings of the Random Forest algorithm. Unlike k-NN it does not require choosing a
339 hyperparameter value like k, rather the inbuilt processes of random feature split criteria, creation of multiple
340 decision trees and cumulating the intermediary results from these trees results in the observed outstanding
341 performance. The diverse trees used in the decision-making process bolster the accuracy and stability of the final
342 prediction. The final numeric metric results and AUC for Classifier 5 are listed in the following table and graph
343 **(Table 6, Fig. 8).**

Table 6: Classifier 5 Experiment Results for the Random Forest with 35 cytokines

Algorithm	Classification Criterion	Predictor Feature Space	AUROC	Prediction Accuracy	Sensitivity	Specificity
Random Forest	Decision Trees	35 Cytokines	**.99**	.96	.954	.967

344

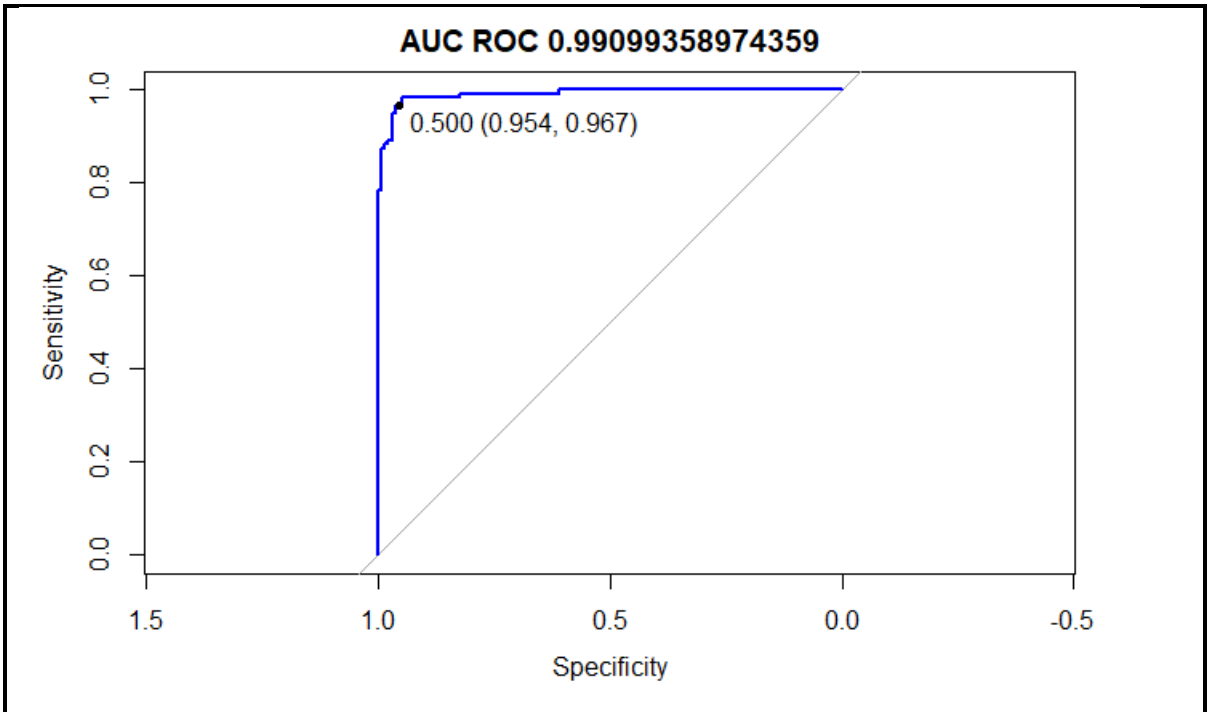

Fig. 8: Classifier 5 Experiment: ROC-AUC curve for Random Forest with the total set of 35 cytokines.

345 ## Discussion

346 Random Forest with 35 cytokines (**Classifier 5 Experiment**) overall proved to be the best ML classification
347 technique as it demonstrated a AUROC of **.99**. The Random Forest out-performed other models and demonstrated
348 an almost perfect AUCROC because it is composed of an ensemble of uncorrelated decision tree models that

collectively provide a better classification compared to that generated by individual models. Random Forest like bagging creates trees from bootstrap samples. Additionally, Random Forest selects a subset of features at each partition that creates trees. This is a desirable characteristic resulting diversity within the tree predictions and uncorrelated prediction errors. A random subset of features is used at each split point to create trees and this diversifies the ensemble therefore enabling better overall performance than algorithms like k-NN.

k-NN with 10 cytokines (**Classifier 3 Experiment**) selected by the Random Forest variable importance mechanism provided the next best AUROC value of **.96**. The variable importance of cytokines was ascertained by the Gini index , a measure that ascertains how each feature contributes to the homogeneity of the nodes and leaves in the resulting Random Forest. Each time a feature is used to split a node the Gini impurity index of the child nodes is calculated and compared to the original node.

Classifier 1(**AUROC=.95**),2**(AUROC=.93**) and 4(**AUROC=.92**) provided reasonably good accuracy but were not as effective as Classifier 3(**AUROC=.96**) and 5(**AUROC=.99**).

Conclusions

This research uniquely implements the use of cytokine plasma biomarkers to differentiate CAD from Control cases. Additionally, it emphasizes the exploratory paradigm of multiple classifier experiments that show improved prediction accuracy across different models. The k-NN algorithm implementations were compared in terms of efficacy as well as juxtaposed relative to the performance of Random Forest algorithm. Overall, the use of Feature importance for cytokine selection improved the prediction accuracy across k-NN algorithmic experiments since this distinction eliminates noisy data and identifies the key biomarkers used for the final classification. The insignificant cytokines degraded the performance of k-NN which is sensitive to noise while Random Forest was minimally affected. As compared to prior research studies [6], [13] that used Random Forest with cytokines to differentiate disease groups from controls, our study exhibited better AUROC **(.99**) .

The current research prediction accuracy for noisy data obtained using different k-NN experimental setups was better than the methodology employed in the diabetes case study [14] .The Random Forest implementation specifically with the total set of 35 cytokines provided .99 AUROC outperforming the previous research studies [29, 30].

Overall, both Random Forest and k-NN generated reasonably good results with all the cytokines and cytokines selected as per their importance determined by Random Forest, ReliefF and Boruta. For both k-NN and Random Forest classifier experiments were balanced for the bias variance trade off by performing cross-validation, data augmentation data balancing and using the 75%-25% split towards training and testing set. Overall Random Forest provided superior AUROC (**.99**) and prediction accuracy (**.96**) metrics.

In this age of innovation and all-pervasive ML systems, it is important to leverage the abstraction, generalization, optimization, and computational power of the versatile ML algorithms that can be used across a wide spectrum of domains of which medical sciences is a very prominent one. Future investigative research of the role of cytokines for identifying inflammation suffered by CAD subjects will translate to therapeutic targets.

Contemporary research indicates that numerous biological factors contribute to risk of CAD including individual molecular species of lipoproteins, oxidative stress, and genetic determinants of inflammation and coagulopathy, among others. The analytical mathematical techniques emerging from this research will permit the analysis of a much broader array of factors, including their mutual interaction in the appreciation of risk of CAD. The inclusion of a broad array of cytokines will contribute a new dimension to this analysis that can lead to improved risk prediction and novel therapeutic interventions.

Ethics approval and consent to participate

This study was approved by the UCSF Institutional Review Board Committee on Human Research and conducted in accordance with the principles of the Declaration of Helsinki, and all subjects provided written informed consent prior to participation.

Consent for publication

Not applicable

Availability of data and materials

The data used for this research comprises confidential patient health information. The data obtained from the Kane laboratory and Genomic Resource in Cardiovascular and Metabolic Disease at UCSF are HIPPA protected and cannot be released publicly. The data, without personal identification, are delineated within the manuscript and we are prepared to answer any questions that researchers might have regarding the data usage for the experimental framework.

Abbreviations

AUROC: Area under Receiver Operating Characteristic

CAD: Coronary Artery Disease

FPR: False Positive Rate

k-NN: K Nearest Neighbor

ML: Machine Learning

ROSE: Random Over-sampling Examples

TPR: True Positive Rate

Competing interests

The authors report no conflicts of interest.

Funding

This research was supported by the NIH under Ruth L. Kirschstein National Research Service Award 2T32HL007731-26 from the Department of Health and Human Services Public Health Services (KTC). Additional support was provided by the Read Foundation Charitable Trust and the Campini Foundation (JPK).

Authors' contributions

SEEMA conceived the data mining discovery plan, implemented the Machine learning algorithm, and wrote the paper. PANKAJ provided valuable advice with regards to Statistical and Machine Learning technical details. KATE conceptualized, supervised, and performed lab experiments, and provided the relevant data details. JAMES conducted the lab experiments and helped coordinate the research collaboration. EVELINE, MARY, and JOHN collected the clinical data and provided expertise from the biomedical perspective to direct the experimental research framework that resulted in extracting the final insights.

All authors helped with the analysis and preparation of the Manuscript. They have read and approved the final manuscript.

Acknowledgements

Not applicable.

References

[1] ""Cardiovascular Diseases (CVDs)." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).".

[2] "Namara, Kevin Mc, et al. "Cardiovascular Disease as a Leading Cause of Death: How Are Pharmacists Getting Involved?" Integrated Pharmacy Research and Practice, Volume 8, 2019, pp. 1–11., doi:10.2147/iprp.s133088.".

[3] "Hastie, Trevor, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2017.".

[4] "Zhang, Jun-Ming, and Jianxiong An. "Cytokines, Inflammation, and Pain." International Anesthesiology Clinics, vol. 45, no. 2, 2007, pp. 27–37., doi:10.1097/aia.0b013e318034194e.".

[5] "Dinarello, Charles A. "Historical Insights into Cytokines." European Journal of Immunology, U.S. National Library of Medicine, Nov. 2007, www.ncbi.nlm.nih.gov/pmc/articles/PMC3140102/.".

[6] "Yu, Linghua, et al. “Inflammatory Profiles Revealed the Dysregulation of Cytokines in Adult Patients of HFMD.” International Journal of Infectious Diseases, vol. 79, 2019, pp. 12–20., doi:10.1016/j.ijid.2018.11.001.".

[7] "Thompson, Peter L., and S. Mark Nidorf. “Anti-Inflammatory Therapy with Canakinumab for Atherosclerotic Disease: Lessons from the CANTOS Trial.” Journal of Thoracic Disease, vol. 10, no. 2, 2018, pp. 695–698., doi:10.21037/jtd.2018.01.119.".

[8] "Creasy, Kate Townsend, et al. “Abstract 20918: Cytokines Involved in Arterial Wall Inflammation Are Transported by High Density Lipoprotein Particles.” Circulation, 9 June 2018, ahajournals.org/doi/10.1161/circ.136.suppl_1.20918.".

[9] R. e. a. “. C. A. D. v. D. M. A. b. C. L. a. E. F. R. i. C. M. K. A. 2. w. Alizadehsani, "Alizadehsani, Roohallah, et al. “Diagnosing Coronary Artery Disease via Data Mining Algorithms by Considering Laboratory and Echocardiography Features.” Research in Cardiovascular Medicine, Kowsar, Aug. 2013, www.ncbi.nlm.nih.gov/pubmed/25478509.".

[10] Q.-u.-a. e. a. “. D. o. C. A. D. A. R. a. W. C. R. a. P. H. 4. F. 2. w. Mastoi, "Mastoi, Qurat-ul-ain, et al. “Automated Diagnosis of Coronary Artery Disease: A Review and Workflow.” Cardiology Research and Practice, Hindawi, 4 Feb. 2018, www.hindawi.com/journals/crp/2018/2016282/.".

[11] N. e. a. “. L. f. A. o. C. A. D. i. C. C. A. S. F. i. C. M. F. M. S. 2. N. 2. w. Hampe, "Hampe, Nils, et al. “Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey.” Frontiers in Cardiovascular Medicine, Frontiers Media S.A., 26 Nov. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6988816/.".

[12] C. e. a. “.-B. C. D. W. M. L. A. R. F. F. 6. J. 2. w. Martin-Isla, "Martin-Isla, Carlos, et al. “Image-Based Cardiac Diagnosis With Machine Learning: A Review.” Frontiers, Frontiers, 6 Jan. 2020, www.frontiersin.org/articles/10.3389/fcvm.2020.00001/full.".

[13] "Struck, Nicole S, et al. “Cytokine Profile Distinguishes Children With Plasmodium Falciparum Malaria From Those With Bacterial Blood Stream Infections.” The Journal of Infectious Diseases, vol. 221, no. 7, 2019, pp. 1098–1106., doi:10.1093/infdis/jiz587.".

[14] "Kandhasamy, J. Pradeep, and S. Balamurali. “Performance Analysis of Classifier Models to Predict Diabetes Mellitus.” Procedia Computer Science, vol. 47, 2015, pp. 45–51., doi:10.1016/j.procs.2015.03.182.".

[15] "Jabbar, M. Akhil, et al. “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.” Procedia Technology, vol. 10, 2013, pp. 85–94., doi:10.1016/j.protcy.2013.12.340.".

[16] "Enriko, I Ketut & Suryanegara, Muhammad & Gunawan, Dinda. (2016). Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters. 8. 59-65.".

[17] "Faizal, Edi, and Hamdani Hamdani. “Weighted Minkowski Similarity Method with CBR for Diagnosing Cardiovascular Disease.” International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018, doi:10.14569/ijacsa.2018.091244.".

[18] "Saini, Indu, et al. “QRS Detection Using K-Nearest Neighbor Algorithm (KNN) and Evaluation on Standard ECG Databases.” Journal of Advanced Research, vol. 4, no. 4, 2013, pp. 331–344., doi:10.1016/j.jare.2012.05.007.".

[19] P. M. e. a. “. T. w. C. f. A. D. N. N. E. J. o. M. 2. S. 2. w. Ridker, "Ridker, Paul M, et al. “Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease: NEJM.” New England Journal of Medicine, 21 Sept. 2017, www.nejm.org/doi/10.1056/NEJMoa1707914.".

[20] "Dinarello, Charles A. “Overview of the IL-1 Family in Innate Inflammation and Acquired Immunity.” Immunological Reviews, U.S. National Library of Medicine, Jan. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5756628/.".

[21] "Stone, M. “Cross-Validatory Choice and Assessment of Statistical Predictions.” Journal of the Royal Statistical Society: Series B (Methodological), vol. 36, no. 2, 1974, pp. 111–133., doi:10.1111/j.2517-6161.1974.tb00994.x.".

[22] "Iyer, Shankar Subramanian, and Gehong Cheng. “Role of Interleukin 10 Transcriptional Regulation in Inflammation and Autoimmune Disease.” Critical Reviews in Immunology, U.S. National Library of Medicine, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC341".

[23] "Fix, E. and Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1951.".

[24] "Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26. doi:http://dx.doi.org/10.18637/jss.v028.i05".

[25] "Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.".

[26] (. U. V. I. i. F. o. .. www.researchgate.net/publication/264046801_Understanding_variable_importances_in_Forests_of_randomized_trees., "(PDF) Understanding Variable Importances in Forests of ... www.researchgate.net/publication/264046801_Understanding_variable_importances_in_Forests_of_randomized_trees.".

[27] "Urbanowicz, Ryan J., et al. “Relief-Based Feature Selection: Introduction and Review.” Journal of Biomedical Informatics, Academic Press, 18 July 2018, www.sciencedirect.com/science/article/pii/S1532046418301400.".

[28] M. B. a. W. R. R. “. S. w. t. B. P. J. o. S. S. w. Kursa, "Kursa, Miron B., and Witold R. Rudnicki. “Feature Selection with the Boruta Package.” Journal of Statistical Software, www.jstatsoft.org/article/view/v036i11.".

[29] "Suvarna, Malini, and Mr.venkategowda N. “Performance Measure and Efficiency of Chemical Skin Burn Classification Using KNN Method.” Procedia Computer Science, vol. 70, 2015, pp. 48–54., doi:10.1016/j.procs.2015.10.028.".

[30] W. H. O. w.-r.-s.-d.-(. “Cardiovascular Diseases (CVDs).” World Health Organization, ""Cardiovascular Diseases (CVDs).” World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).".

429

430 **********