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 current research took an innovative approach to implement K Nearest Neighbor (k-NN) and ensemble Random Forest Machine Learning algorithms to achieve a targeted "At Risk" Coronary Artery Disease (CAD) classification. To ensure better generalizability mechanisms like k-fold cross validation, hyperparameter tuning and statistical significance ($p<.05$) were employed. The classification is also unique from the aspect of incorporating 35 cytokines as biomarkers within the predictive feature space of Machine Learning algorithms.

Results

A total of seven classifiers were developed, with four built using 35 cytokine predictive features and three built using 9 cytokines statistically significant ($p<.05$) across CAD versus Control groups determined by independent two sample t tests. The best prediction accuracy of 100% was achieved by Random Forest ensemble using nine significant cytokines. Significant cytokines were selected to decrease the noise level of the data, allowing for better classification.

Additionally, from the bio-medical perspective, it was enlightening to empirically observe the interplay of the cytokines. Compared to Controls, moderately correlated (correlation coefficient $r=.5$) cytokines "IL1-β", "IL-10"

[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

were both significant and down regulated in the CAD group. Both cytokines were primarily responsible for the Random forest generated 100% classification. In conjunction with Machine Learning (ML) algorithms, the traditional statistical techniques like correlation and t tests were leveraged to obtain insights that brought forth a role for cytokines in the investigation of CAD risk.

Conclusions

Presently, as large-scale efforts are gaining momentum to enable early detection of individuals at risk for CAD by the application of novel and powerful ML algorithms, detection can be further improved by incorporating additional biomarkers. Investigation of 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 approaches.

Keywords

CAD, Machine Learning, k-NN, Random Forest, Predictive Accuracy, Distance Metrics, k fold Cross Validation, Classification, Plasma cytokines.

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 Machine Learning algorithms like k-NN, 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 statistical significance in conjunction with k-fold cross validation, a powerful resampling technique that overcomes the issue of overfitting, ensuring better generalizability of the model [3]. Cytokines were considered significant by an independent two samples t test ($p<.05$) between CAD versus Control groups.

We used plasma cytokines as novel biomarkers to improve classification in patients with or without clinical coronary disease. This approach promises to identify mechanisms of disease, 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 resulting in pro-inflammatory or anti-inflammatory response. 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. 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 has been discussed here. Random Forest [3], an ensemble algorithm though not as interpretable as k-NN is also a very innovative technique that has been empirically proven to generate high predictive accuracy.

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 AUC value of 91, indicative of its excellent partition efficacy.

Struck, et al. [9] 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 Machine Learning 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 directives to develop new point-of-care tests in Sub-Saharan Africa.

Kandhasamy et al [10] tackles the early prediction of diabetes, a commonly occurring disease in the contemporary context using important machine learning 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. [11] 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 [12] experimented and compared Machine Learning 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 used were chosen because they were simple to measure and provided a better prediction accuracy in the context of k-NN compared to Naïve Bayes and Decision Trees. The usage of multiple algorithms provided a comparative analysis with regards to evaluation measurement like prediction accuracy.

Faizal, Edi, and Hamdani Hamdani. [13] extracted the early diagnosis of CAD based on feature similarity with predictive attributes like age, gender, and risk factor 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 was less than .80, the diagnosis would be determined with the consultation of an expert. The performance metrics like Prediction Accuracy and Sensitivity obtained by the application of this methodology were high.

Saini, Indu, et al. [14] 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 current study endeavored to improve and extend the techniques reviewed in the aforementioned studies. The objective of minimizing and downplaying noisy data were the motivation behind the current research's usage of significant cytokines as well as that of weighted k-NN. The usage and comparison of multiple algorithms has proven an effective way to obtain a holistic view with regards to classification. The current research incorporates this comparison paradigm to present the results. The empirical proof of enhanced performance by the usage of hyperparameter tuning and cross fold validation [15] for Random Forest and k-NN directed the in-depth exploration conducted in the current research effort. Fine tuning hyperparameters in general has proven to be an effective optimization technique.

Methods

The data set is composed of biomarker levels for 104 individuals. Thirty-five cytokine biomarkers were measured for each individual in addition to the final target feature attribute which categorizes whether the individual belongs 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 females subjects, ages 18 to 65 (median age = 42) with diagnosed CAD and age, 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 diagnosis, 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 quantification raw data were analyzed by xPonent software and were expressed in pg/mL using the standard curve for each cytokine assayed. The following table(**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

The detailed statistical summaries and statistical significance of 35 cytokines, CAD versus Control is provided in the following (**Table 2**) table.

Table 2: 35 Cytokines predictor biomarkers summaries and p-value to differentiate CAD versus Control in terms of significance. The summary descriptions included mean, standard deviation, minimum, maximum, first Quartile and third quartile.

Cytokine Predictor Feature	Mean	Standard Deviation	Minimum	25th Percentile	75th Percentile	Maximum	Significance p-value (<.05)
FGFbasic	11.996	9.635	1.955	7.671	13.742	72.050	0.4240
IL-1β	5.771	3.016	1.288	2.160	8.088	9.258	0.0094*
GCSF	45.633	25.582	5.881	27.239	63.091	145.104	0.6483
IL-10	20.219	15.431	0.071	6.550	26.970	105.783	0.0017*
IL-13	7.174	4.204	0.819	3.727	9.230	30.103	0.3990
IL-6	5.670	3.460	0.176	2.677	7.263	22.038	0.7391
IL-12	66.702	41.517	22.277	42.731	79.971	350.602	0.0364*
RANTES	1,771.252	736.181	65.956	1,303.297	2,300.639	3,408.323	0.9931
Eotaxin	15.987	11.115	4.141	8.110	20.635	72.468	0.0063*
IL-17A	4.467	3.538	0.954	1.600	6.650	26.940	0.4287
MIP1-α	27.833	11.566	7.122	19.109	32.160	93.781	0.1687
GM-CSF	11.424	7.093	1.319	3.209	16.788	23.168	0.0048*
MIP1-β	38.695	40.365	11.646	24.734	42.984	420.721	0.5060

MCP-1	124.139	70.882	52.538	75.558	148.525	473.368	0.1526
IL-15	56.010	71.060	5.874	15.203	65.586	580.576	0.6764
EGF	10.255	8.299	0.216	2.812	12.477	69.896	0.9185
IL-5	5.079	4.400	0.079	1.555	6.654	35.268	0.0527
HGF	163.913	157.597	20.805	116.986	165.058	1,472.905	0.8866
VEGF	1.590	1.057	0.045	0.627	2.067	8.423	0.3923
IL-1α	5.144	3.218	0.588	2.749	7.216	25.991	0.0166*
IFN-γ	5.824	1.446	1.202	5.106	6.661	9.237	0.0664
IL_17F	87.508	95.162	0.144	22.817	99.818	734.199	0.5680
IFN-α	16.572	16.087	1.715	9.785	15.863	95.497	0.4661
IL_9	2.866	3.520	0.346	1.790	3.332	35.652	0.7874
IL_1RA	56.313	138.716	8.080	15.011	43.673	1,326.906	0.3164
TNF-α	7.727	4.685	0.074	2.572	11.112	20.975	0.0221*
IL-3	8.549	7.705	0.996	3.246	12.676	69.433	0.1243
IL-2	7.265	4.883	0.041	2.218	9.541	30.752	0.0008*
IL-7	3.011	3.715	0.495	2.012	3.017	34.318	0.2807
IP-10	15.616	14.559	3.211	8.322	16.524	90.918	0.7864
IL-2R	71.827	38.244	1.084	53.837	89.784	288.537	0.9605
IL-22	22.518	38.360	4.279	4.801	24.750	260.526	0.7506
MIG	29.632	60.289	2.261	13.424	29.606	596.227	0.72822
IL-4	42.553	27.117	2.511	8.262	62.744	81.863	0.0036*
IL-8	12.810	9.272	0.145	6.219	17.451	81.982	0.6407

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. Finally, fifty percent of the data were randomly assigned to the

training and the rest to the test set to eliminate selection bias. Given the small data set (n=104), the 50% split was decided upon after a series of test runs to achieve the best prediction accuracy

k-NN K-Nearest Neighbor

k-NN, a supervised ML algorithm was initially proposed by Fix and Hodges [16]. It is based upon the similarity paradigm, indicating that the classification of unlabeled examples are differentiated by means of distance metrics and are finally ascribed the class of k($k \geq 1$) nearest neighbors. k-NN, by the 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 [12] 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. k-NN Algorithmic Steps

Step 1: Read the data and assign k(number of neighbors).

Step 2: Conduct steps 3-5 until the entire test data has been classified.

Step 3: Iterate through the training set to obtain the predicted class / class label, then compute the distance between each instance of test data and each row of the training data.

Step 4: Sort the distances in ascending order and choose the first k distances.

Step 5: Choose the most frequently occurring class from these k distances and assign it to the test data.

Step 6*:* Return the class of the test data.

2. 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 weightage by the process of majority vote.

3. 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, weighted Minokwski distance, Manhattan distance, Canberra distance, Chebyshev distance and the Cosine distance. Much of the current analysis was implemented using the caret R Package [17], Canberra R package and ggplot2 [18]package.

The following distances were implemented for the current research:

3.1 Minokwski 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:

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

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

Weighted Minkowski distance is used to identify the importance of a feature attribute in context of the overall classification.

To incorporate the weightage, each feature is assigned a weighing coefficient w_i (i=1……..D) and therefore the weighted Minkowski distance function is mathematically represented as follows:

$$\mathbf{d_w(X, Y) = (\sum_{i=1}^{D} w_i |x_i - y_i|^p)^{1/p}} \quad (\mathbf{2})$$

3.2 Euclidean distance

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

$$\mathbf{d(x, y) = ||x - y||_2 = (\sum_i^D |x_i - y_i|^2)^{1/2}} \quad (\mathbf{3})$$

If p=1, L_1-norm is the Manhattan distance.

$$\mathbf{d(x, y) = ||x - y||_1 = \sum_i^D |x_i - y_i|} \quad (\mathbf{4})$$

3.3 Canberra Distance

Canberra is an algorithm that represents the weighted version of the Manhattan distance. This metric specifies the distance between pair of points in the vector space. This distance measure is ascertained by computing the absolute value of the difference between pair of vector points divided by the sum of the absolute values of the

vector points. This distance is used primarily for values that are represented in a multidimensional space as well as a dataset whose values vary to a small extent close to zero.

Mathematically, the Canberra distance d between vectors x_i and y_i in a D-dimensional real vector space is given by the following formula:

$$\mathbf{d}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{i}}^{\mathbf{D}} \frac{|\mathbf{x_i} - \mathbf{y_i}|}{|\mathbf{x_i}| + |\mathbf{y_i}|} \qquad (\mathbf{5})$$

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 get 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 a 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 with regards to training data, but also with regards to node split feature partition.

The steps entailing the implementation of Random Forest algorithm are as follows:

Step 1: Read the data and select random samples.

Step 2: Conduct steps 3-4 until the entire test data gets classified.

Step 3: Create a decision tree for each sample by random partition split feature criterion. Generate a classification result for each decision tree.

Step 5: Perform majority voting for every classified result.

Step 6: Majority vote is applied to assign the final class label.

Step 7 Return the class of the test data.

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 then finally validating it on one partition. This study used k=10 to implement 10-fold cross validation.

Following steps delineate the process of k-fold cross-validation.

Step 1: Read the data and randomize it.

Step 2: Partition the dataset into k-1 groups for training and one group for validation.

Step 3: Iterate through the k-1 training group to fit a model with a set of tuning parameters.

Step 4: Evaluate the validation models to determine the best tuning parameter for optimizing classification accuracy or prediction by averaging to obtain the model's potential generalizability skill.

Step 5: Finally, the test is run to provide a realistic classification or prediction using the optimal tuning parameters.

2. Significant Cytokines

The investigation also entails narrowing down the feature space, using the attributes of paramount importance by selecting the ones significantly different in the CAD versus Control groups. The independent two samples t-test was used at a threshold significance level(α) =.05.

Classifier Experimental Framework

The first classifier experiment entailed applying the k-NN algorithm involving 35 cytokine predictor features with the Euclidean distance and 10-fold cross-validation. The second classifier experiment involved the usage of k-NN algorithm incorporating only significant cytokines as predictor features with the Euclidean distance and 10-fold cross-validation. The third classifier experiment evaluated the usage of 10-fold cross-validated, and weighted k-NN with all cytokines across a variety of kernel functions like inverse, rectangular and triangular etc. The fourth classifier experiment implemented 10-fold cross-validated, and weighted k-NN with kernel function and only significant cytokines. The fifth classifier experiment implemented Random Forest using the 35 cytokines with 10-fold cross-validation. The sixth classifier experiment applied Random Forest with only significant cytokines and 10-fold cross-validation. Finally, the seventh classifier experiment entailed using k-NN algorithm with Canberra distance tested across different values of k. For this analysis the predominately R programming packages caret [17], knngarden, corrplot [19] and ggplot2 [18] were used to obtain the results. The 10-fold cross-validation resampling technique was used for all classifier experiments except classifier 10 which entailed k-NN algorithm implemented with the Canberra distance measure.

Evaluation Measures

A versatile set of performance evaluation measures were used to obtain insight related to the efficacy of algorithms in conjunction with achieving the objective of juxtaposing the spectrum of algorithms experimented on. Mainly, the focus was on the following numerical measures:

1. Prediction Accuracy

Fundamentally, Prediction Accuracy is the key metric that accesses the quality of a model in terms of how well the test data is classified as per the targeted feature. Mathematically this can be depicted as follows:

$$\textbf{Prediction Accuracy} = \frac{\textbf{Number of correct predictions}}{\textbf{Total number of Predictions}} \quad (\mathbf{6})$$

In context of the current scenario, where we are dealing with a binary classification of CAD versus Control , it can also be quantified in terms of true positive and true negative measures.

$$\textbf{Prediction Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}} \quad (\mathbf{7})$$

Where, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

2. Sensitivity (TPR: True Positive Rate)

This metric numerically quantifies the performance of a ML algorithm from the perspective of the proportion of actual positive instances that were predicted to be positive. Alternatively, it is also called the True positive rate.

$$\textbf{Sensitivity} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \quad (\mathbf{8})$$

Where, TPR=1-TNR(True Negative Rate)

3. Specificity (TNR: True Negative Rate)

This measure quantifies the proportion of actual negative instances that were predicted to be negatives. Alternatively, it is also called the True negative rate.

$$\textbf{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN + FP}} \quad (\mathbf{9})$$

Where TNR =1-TPR(True Positive Rate)

4. AUC-ROC (Area Under the Curve-Receiver Operating Characteristics)

This is a vital measure that helps determine the degree of separability achieved by the relevant Classification algorithm. Higher AUC increases the algorithm's capability of differentiating into classes. The ROC curve is created by plotting FPR (x-axis) against the TPR (y-axis).

5. T tests for Statistical Significance

To improve the efficacy of classification and for optimal feature selection, independent two sample t tests were used to extract the features that are significantly different for CAD versus Control plasma cytokine levels. The significance level (α) of .05 was used to differentiate the statistical significance versus the insignificance.

6. Correlation Coefficient Matrix

A correlation coefficient matrix was constructed to ascertain the strength, direction, and the statistical significance of pairwise association pertaining to cytokines that were mainly responsible for the classification of at-risk CAD group versus the not at-risk Control group. The Pearson correlation coefficient matrix provided a better understanding of the relational bio marker underpinnings. To identify significance of the correlational association, independent two sample t tests were used with a threshold significance level of α=.05.

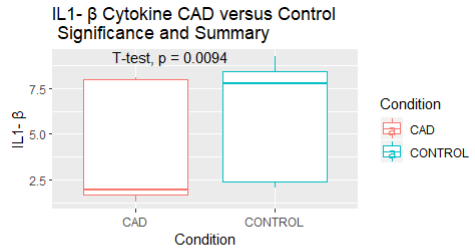
Results

Identification of Significant Cytokines for CAD classification

The following graphs and tables provide a comparative analysis of significant cytokines grouped by CAD versus Control, with the independent two samples t test determining significance. Specifically, the significant cytokines were “IL1-β”, “IL-10”, “IL-12”, “Eotaxin”, “GM-CSF”, “IL1-α”, “TNFα”,“IL-2” and “IL-4”.

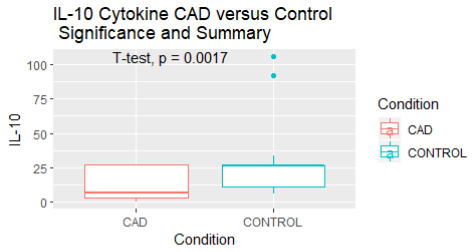
The following Figure 1(**Fig. 1 A-I**) displays boxplot ,numerical summaries differentiated by CAD versus Control classification and comparative cytokine patterns with baseline CAD measures (**Fig. 1 J**).

A



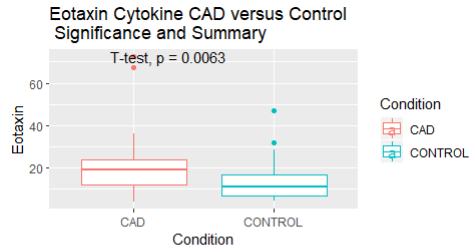
Condition	length	mean	sd
CAD	39	4.743005	3.215707
CONTROL	65	6.388524	2.732986

B



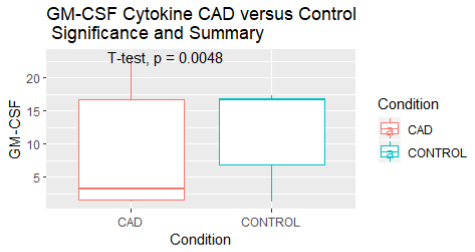
Condition	length	mean	sd
CAD	39	14.53343	12.28508
CONTROL	65	23.63114	16.18836

C



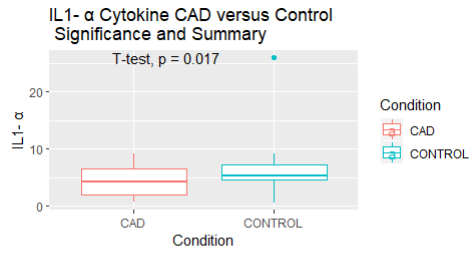
Condition	length	mean	sd
CAD	39	20.35619	14.084731
CONTROL	65	13.36498	7.900718

D



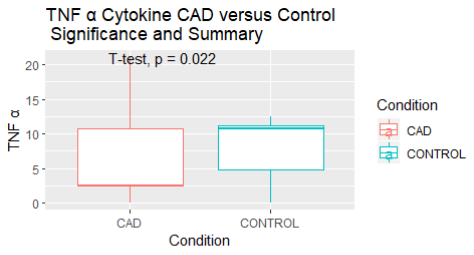
Condition	length	mean	sd
CAD	39	8.770445	7.684905
CONTROL	65	13.016063	6.247694

E



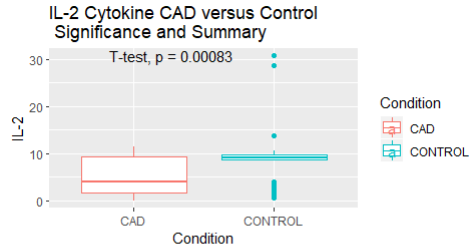
Condition	length	mean	sd
CAD	39	4.242949	2.534670
CONTROL	65	5.685304	3.472518

F



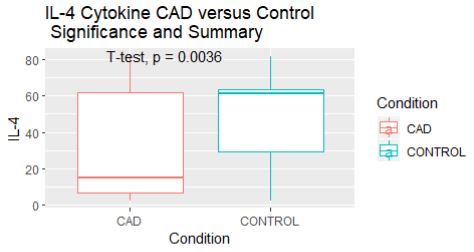
Condition	length	mean	sd
CAD	39	6.304236	5.112495
CONTROL	65	8.580646	4.222346

G



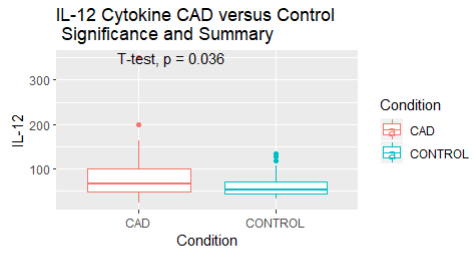
Condition	length	mean	sd
CAD	39	5.334653	4.013490
CONTROL	65	8.423780	5.016765

H



Condition	length	mean	sd
CAD	39	32.36948	27.69523
CONTROL	65	48.66302	25.03735

I



Condition	length	mean	sd
CAD	39	79.98842	58.83928
CONTROL	65	58.73000	23.37330

J

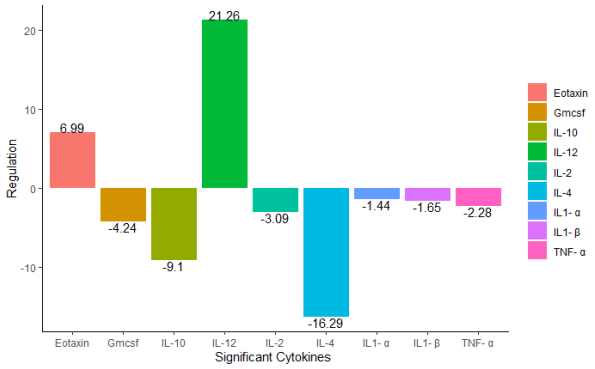


Fig. 1 Significant Cytokine Comparisons using Boxplot, Summaries and Association (CAD versus Control). Significance was ascertained using independent sample t test (p<.05).

A."IL1-β" **B.**"IL-10" **C.**"IL-12" **D.**"Eotaxin" **E.**"GM-CSF" **F.**"IL1-α" **G.**"TNFα" **H.**"IL-4" **I.**"IL-2".

J. Up and down association of CAD versus Control with baseline Control.

Resampling Cross Validation Results

The following classifier experiments, tables and graphs display 10-fold cross validation resampling experimental results (**Tables 1-6, Fig. 2-7**) obtained for each classifier experiment and the simulation results obtained by varying tuning parameter k.

1. Classifier 1 Experiment

The first classifier was implemented using 35 cytokines predictor feature variables with the "At Risk" feature as the target classification variable. k-NN was used as a classification algorithm with the Euclidean distance metric. The highest resampling prediction accuracy obtained was acceptable(.6611111) though not very good and this was achieved for a k value of 1.The resampling results for this classifier have been displayed in the following table and graph(**Table. 3,Fig. 2**)

Table. 3: Classifier 1 Experiment: Ten-fold cross-validation resampling prediction accuracy results for k-NN algorithm with 35 cytokines.

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy with k (number of neighbors)
k-NN	Distance Measure: Euclidean	35 Cytokines	k=1 Prediction Accuracy = .6611111

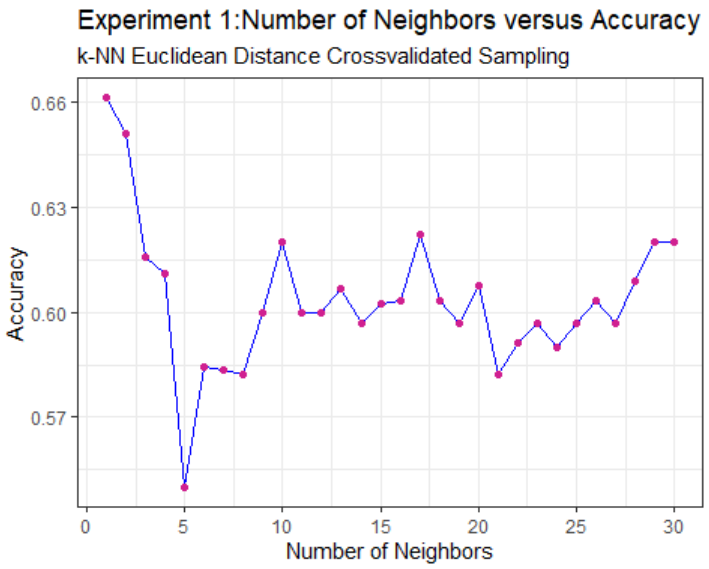


Fig. 2 Classifier 1 Experiment: Ten-fold cross-validation resampling prediction accuracy for k-NN (Euclidean Distance) with 35 cytokines across tuning parameter k (Number of neighbors). The best prediction accuracy of .661111 was obtained for k=1.

2. Classifier 2 Experiment

The second classifier used only 9 significant cytokines based on CAD versus Control, identified using the independent two sample t-test. Like Classifier 1, Classifier 2 also implemented k-NN with the Euclidean distance metric. Classifier 2 out-performed Classifier 1 with its highest resampling prediction accuracy of 0.7612698 for a k value of 5.The improvement can be attributed to the dropping of redundant feature variables. The resampling results for this classifier have been displayed in the following table and graph (**Table. 4,Fig. 3**)

Table. 4: Classifier 2 Experiment: Ten-fold cross-validation resampling prediction accuracy results with 9 significant cytokines and k=5

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy with k (number of neighbors)
k-NN	Distance Measure: Euclidean	9 Significant Cytokines p-value <.05	k=5 Prediction Accuracy = 0.7612698

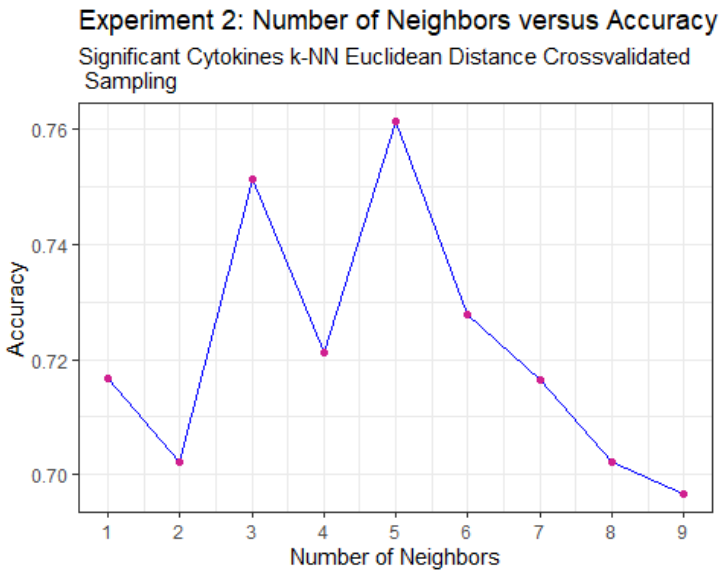


Fig. 3 Classifier 2 Experiment: Ten-fold cross-validation resampling prediction accuracy results with 9 significant cytokines across tuning parameter k (Number of neighbors). The best accuracy of 0.7612698 was obtained for k=5.

3. Classifier 3 Experiment

Classifier 3 used 35 cytokines like Classifier 1 but included the application of k-NN algorithm with weighted Euclidean as the distance metric. The weighted Euclidean distance metric gives precedence to closest neighbors as compared to those that are further apart. The Classifier 3 out-performed Classifier 1, with an accuracy of .6934525 with k=8 but was worse than Classifier 2, where the significant cytokines were implemented. The results of resampling, 10-fold cross validation for Classifier 3 have been displayed below in a table and graph (**Table. 5,Fig. 4**).

Table. 5: Classifier 3 Experiment: Tenfold cross-validation resampling prediction accuracy results for weighted k-NN algorithm with 35 cytokines and k=8.

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy with k (number of neighbors)
k-NN	Distance Measure: weighted Euclidean	35 Cytokines	k=8 Prediction Accuracy =.6934525 distance = 2 and kernel = inverse

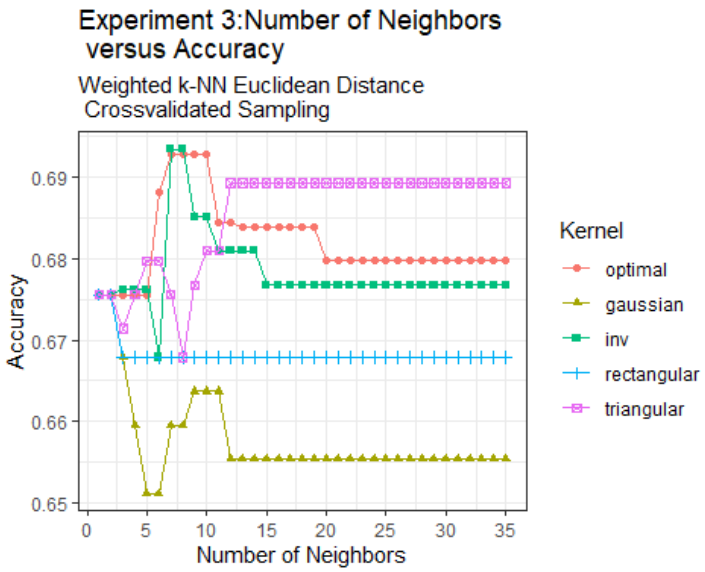


Fig. 4 Classifier 3 Experiment: Ten-fold cross-validation resampling prediction accuracy results for weighted k-NN with 35 cytokines across tuning parameter k (Number of neighbors). The best accuracy of .6934525 was obtained for k=5.

4. Classifier 4 Experiment

This Classifier implemented the k-NN algorithm using weighted Euclidean distance for only 9 significant cytokines. The performance of Classifier 4 with a prediction accuracy of .7446032 for k=6, was better than Classifier 1 and Classifier 3 but not Classifier 2. The improvement can be ascribed to the weighted Euclidean distance in conjunction with the significance of the nine cytokines. The degradation of Classifier 4 performance as compared to Classifier 2 could be due to a small data size. The details related to Classifier 4 has been displayed in the following table and graph (**Table. 6,Fig. 5**).

Table. 6: Classifier 4 Experiment: Ten-fold cross-validation resampling prediction accuracy results for weighted k-NN algorithm with 9 significant cytokines and k=6.

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy with k (number of neighbors)
k-NN	Distance Measure: weighted Euclidean	9 Significant Cytokines	k=6 Prediction Accuracy = 0.7446032 distance = 2 and kernel = optimal

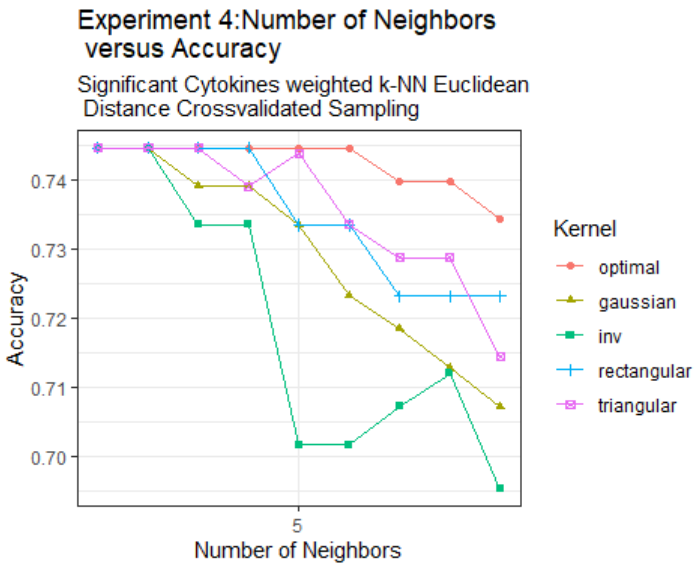


Fig. 5 Classifier 4 Experiment: Ten-fold cross-validation resampling prediction accuracy results for weighted k-NN with 9 significant cytokines across tuning parameter k (Number of neighbors)

5. Classifier 5 Experiment

The Classifier 5 implemented Random Forest using 35 cytokines. The prediction accuracy of .8755556 for Classifier 5 was notably higher than Classifier 1, Classifier 2, and Classifier 3. Random Forest is characterized by implicit partition feature selection. Therefore, the ensemble of decision trees results in a better accuracy by merging the diversity imbued decision trees. The details of the table and graph for classifier 5 are represented below (**Table. 7,Fig. 6**).

Table. 7: Classifier 5 Experiment: Ten-fold cross-validation resampling prediction accuracy results for Random Forest algorithm with 35 cytokines.

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy
k-NN	Random Forest Decision Trees	35 Cytokines	Prediction Accuracy = 0.8755556

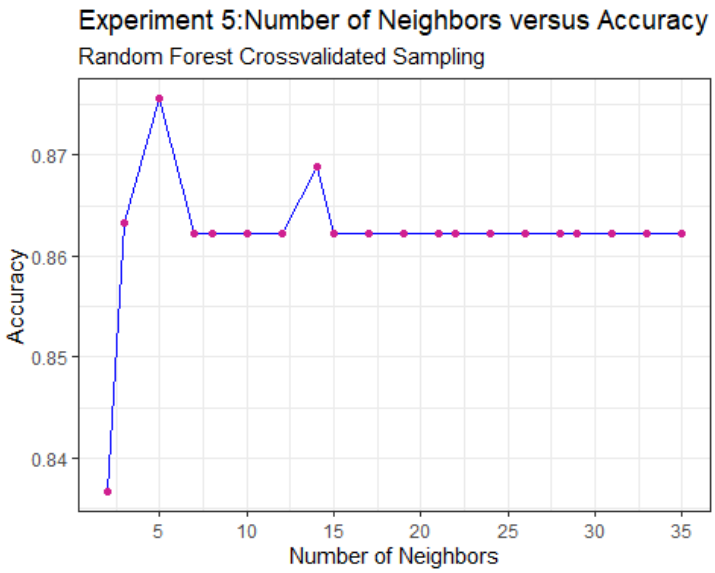


Fig. 6 Classifier 5 Experiment: Ten-fold cross-validation resampling prediction accuracy results for Random Forest with 35 cytokines

6. Classifier 6 Experiment

Classifier 6 was implemented with Random Forest taking only 9 significant cytokines. The resampling 10-fold cross validation prediction accuracy of .9730159 obtained using significant cytokines produced the highest prediction accuracy. The outstandingly high accuracy obtained was due to the usage of significant cytokines in conjunction with the inherent strength of Random Forest. The comprehensive information related to classifier 6 is showcased in the following table and graph. (**Table. 8,Fig. 7**).

Table. 8: Classifier 6 Experiment: Ten-fold cross-validation resampling prediction accuracy results for Random Forest algorithm with 9 significant cytokines.

Algorithm	Classification Criterion	Predictor Feature Space	Optimal Resampling Prediction Accuracy
k-NN	Decision Trees	9 Significant Cytokines p-value <.05	Prediction Accuracy = 0.9730159

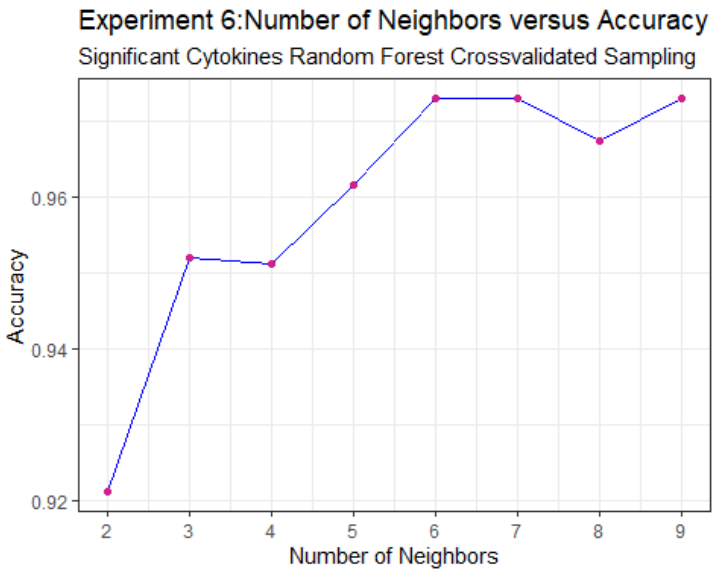


Fig. 7 Classifier 6 Experiment: Ten-fold cross-validation resampling prediction accuracy results for Random Forest with 9 significant cytokines

Testing Results

The testing results obtained by running the algorithm on the test data are displayed in the following tables (**Tables 9-15**) and graphs (**Fig. 8-13**). The testing results applied the hyperparameter tuning metrics optimized by the resampling 10-fold cross validation results. The contribution of significant cytokines with regards to strength of their accurate classification is showcased graphically (**Fig. 14**) . The correlation coefficient matrix for significant cytokines is also displayed graphically (**Fig. 15**).

1. Classifier 1 Experiment

This classifier used k-NN with Euclidean distance measure and k=1 to classify "At Risk" instances. The final prediction accuracy of .7059 was not good but reasonable. For this classification 35 cytokines were used. The extent of separability of CAD versus Control attributed by AUC (.66) was not very remarkable. The details related to numeric measures for Classifier 1 and AUC are provided in the following table and graph. **(Table 9, Fig 8).**

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

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
k-NN	Distance Measure: Euclidean with k=1	35 Cytokines	0.7059	0.5789	0.7812	.66

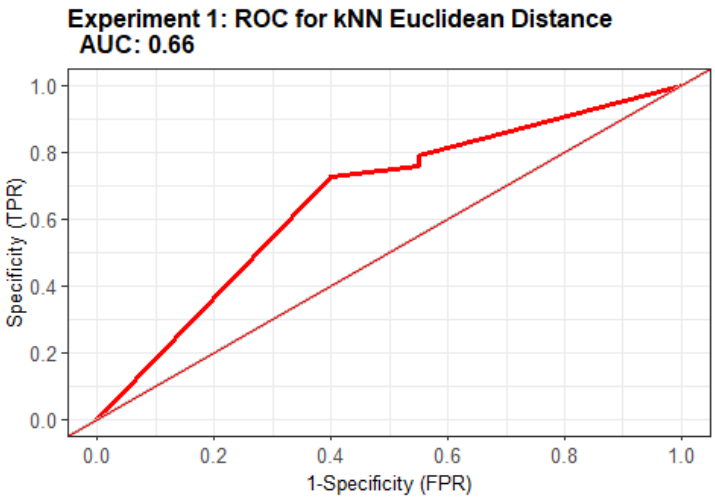


Fig. 8 Classifier 1 Experiment: ROC-AUC curve for k-NN (Euclidean Distance) with 35 cytokines.

2. Classifier 2 Experiment

This classifier used k-NN with Euclidean distance for 9 significant cytokines and k=5. The resulting prediction accuracy of .8293 was better than Classifier 1 for which we used 35 cytokines instead of 9 significant cytokines even though the algorithm used was still k-NN . The enhancement of prediction accuracy can be attributed to the elimination of noisy data by the virtue of dropping non-significant features. AUC value for this classifier was .88, which was much higher than Classifier 1. The numeric results and AUC information are provided in the following table and graph **(Table 10, Fig. 9)**

Table 10: Classifier 2 Experiment Results for the k-NN algorithm with 9 significant cytokines and k=5

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
k-NN	Distance Measure: Euclidean with k=5	9 Significant Cytokines p-value <.05	0.8293	0.6000	0.9615	.88

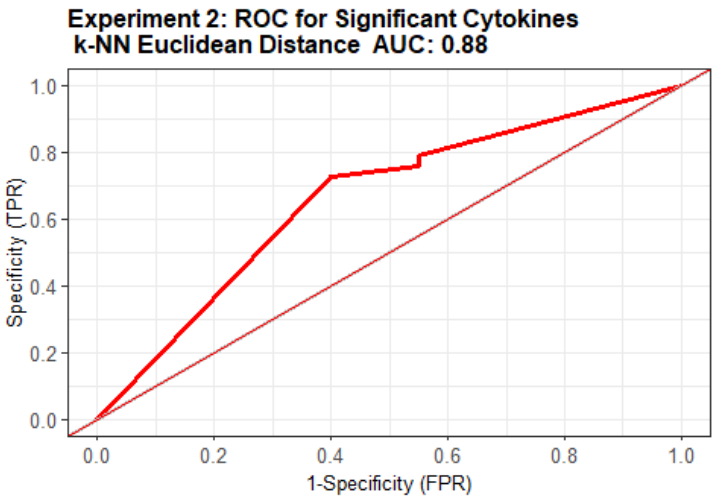


Fig. 9 Classifier 2 Experiment: ROC-AUC curve for k-NN (Euclidean Distance) with 9 significant cytokines.

3. Classifier 3 Experiment

The third classifier applied k-NN with weighted Euclidean distance, inverse kernel and k=8. The experiment incorporated 35 cytokines within the feature space. The prediction accuracy of .72 for this classifier was counterintuitively lower than Classifier 1 and 2.This decrease is probably due to the small data size as well as the noise present due to redundant features used for the purpose. The corresponding AUC was equal to .72,which was also lower than the ones obtained for Classifier 1 and 2.The numerical details and AUC are provide in the following table and graph **(Table 11, Fig. 10)**

Table 11: Classifier 3 Experiment Results for the weighted k-NN algorithm with 35 cytokines and k=8

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
k-NN	Distance Measure: Euclidean Weighted with k=8, kernel = inverse	35 Cytokines	0.72	0.5556	0.8125	.72

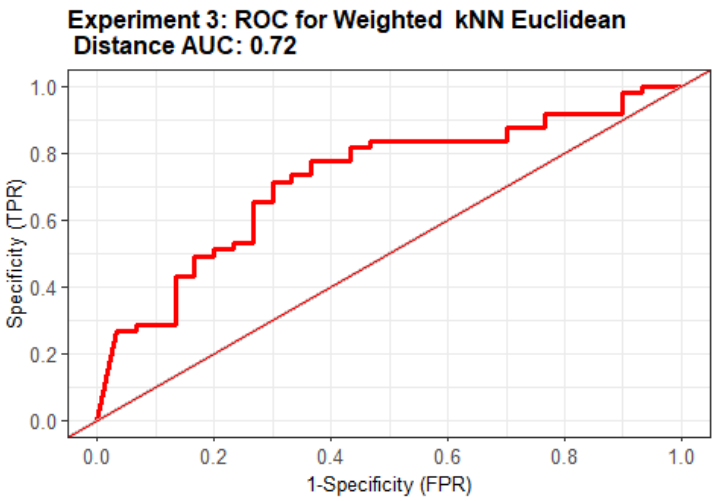


Fig. 10 Classifier 3 Experiment: ROC-AUC curve for weighted k-NN (Euclidean Distance) with 35 cytokines.

4. Classifier 4 Experiment

In contrast to Classifier 3, this classifier used only 9 significant cytokines as predictor features. Like Classifier 3, the algorithm used for this classifier was weighted k-NN although k=5 and optimal kernel was used for this test. As compared to Classifier 3, a prominent augmentation of prediction accuracy was observed for Classifier 4 with its prediction accuracy rising to .878.Classifier 4 achieved the highest prediction accuracy amongst the first four classifiers. The reason for its success is the choice of the most efficacious feature predictors that was ascertained by their significance in terms of CAD versus Control. This result was further enhanced by giving more weight to neighbors that were nearest to the observation instance being classified. The AUC of .78 was achieved, which was less than the AUC for Classifier 2 but more than the AUC for other classifiers discussed thus far. The details pertaining numerical measures and AUC for Classifier 4 is stated in the following table and graph **(Table 12, Fig. 11)**

Table 12: Classifier 4 Experiment Results for the weighted k-NN algorithm with 9 cytokines and k=5

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
k-NN	Distance Measure: Euclidean Weighted with kernel = optimal and k=5	9 Significant Cytokines p-value <.05	0.878	0.7333	0.9615	.78

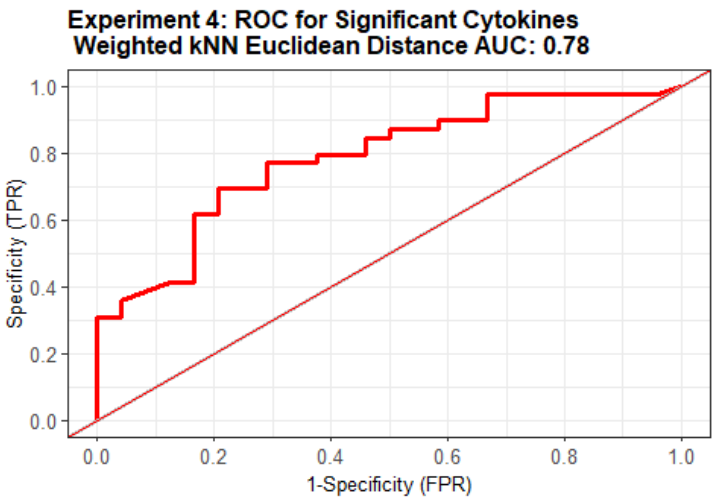


Fig. 11 Classifier 4 Experiment: ROC-AUC curve for weighted k-NN (Euclidean Distance) with 9 significant cytokines.

5. Classifier 5 Experiment

The Classifier 5 implemented Random Forest using 35 cytokines as predictor features. The prediction accuracy of .8824 was extremely high and surpassed the accuracy achieved by the previous four classifiers. This accomplishment can be accounted for by the underpinnings of the Random Forest algorithm. Unlike k-NN it does not require choosing a hyperparameter value like k, inbuilt the process of random feature split criteria, creation of multiple decision trees and cumulating the intermediary results from these trees results in outstanding performance. The diverse trees used in the decision-making process bolsters the accuracy and stability of the final prediction. The numerical value of .96 for AUC was also the highest obtained as compared to the previous four Classifiers. The final numeric metric results and AUC for Classifier 5 are listed in the following table and graph **(Table 13, Fig. 12).**

Table 13: Classifier 5 Experiment Results for the Random Forest algorithm with 35 cytokines.

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
Random Forest	Decision Tree	35 Cytokines	0.8824	0.7368	0.9688	.96

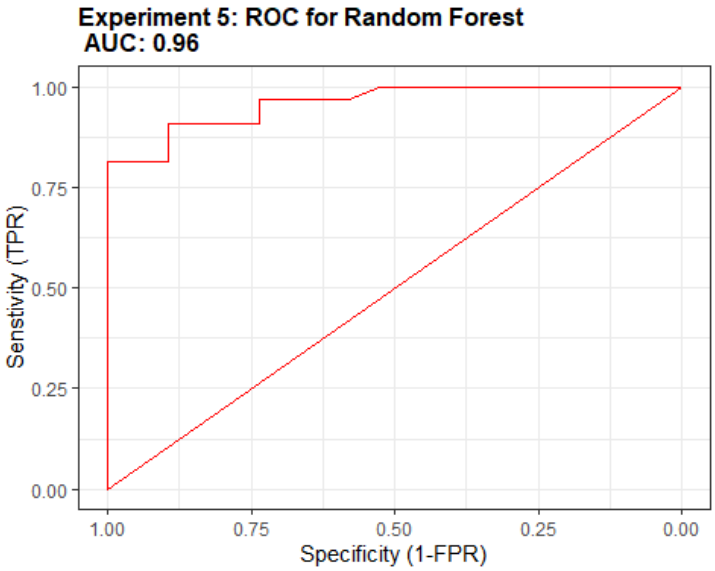


Fig. 12 Classifier 5 Experiment: ROC-AUC curve for Random Forest with 35 cytokines

6. Classifier 6 Experiment

This classifier is like Classifier 5 with regards to algorithm usage except that it used only significant predictor features within its fold. A perfect 100% prediction accuracy was achieved which was not obtained by any of the previous five classifiers. This not only validates the strength of the Random forest algorithm but also highlights the importance of using only the predictor features that contribute the most toward the classification of "At Risk" individuals. Specificity, Sensitivity and AUC were also augmented to 100%. These results are compiled and displayed in the following table **(Table 14, Fig. 13).**

Table 14: Classifier 6 Experiment Results for the Random Forest algorithm with 9 significant cytokines.

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity	AUC
Random Forest	Decision Tree	9 Significant Cytokines p-value <.05	1.000	1.000	1.000	1.00

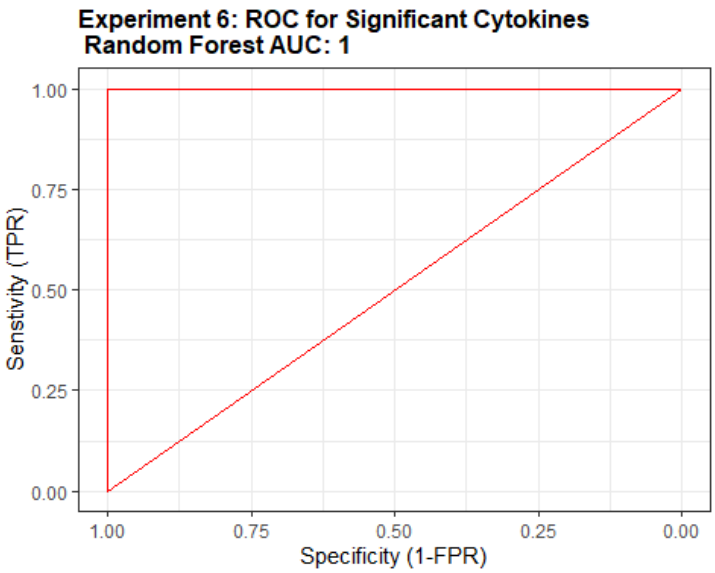


Fig. 13 Classifier 6 Experiment: ROC-AUC curve for Random Forest with 9 significant cytokines.

7. Classifier 6 Significant Cytokine Contribution

The individual contribution of each significant cytokines for the purpose of classification using Random Forest was determined to better understand of the role and impact of cytokines in the risk classification of CAD. The highest contributors for the classification were IL-10 AND IL1-β.The contributions have been displayed in the following table. **(Fig. 14).**

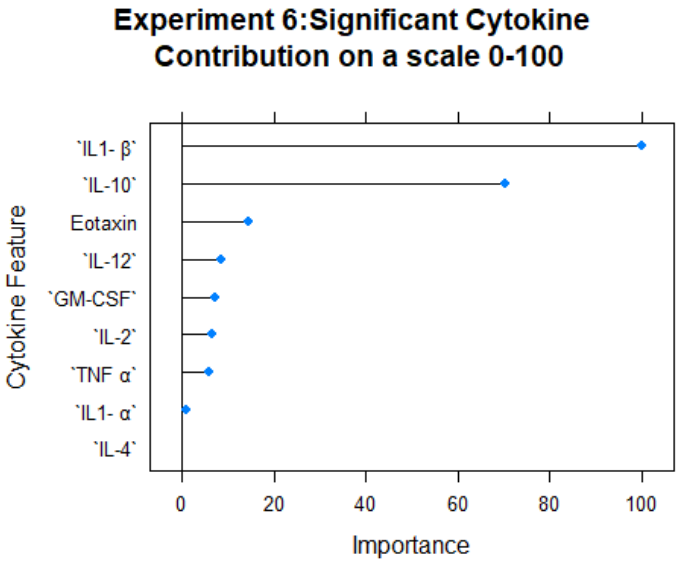


Fig. 14 Classifier 6 Experiment: Cytokine classification contribution for Random Forest with 9 significant cytokines.

8. Significant Cytokine Correlation Matrix

To decipher not only the role of individual cytokines in CAD risk, it was considered imperative to obtain insight related to the interaction between the cytokines themselves which will potentially affect the final classification. The Pearson correlation coefficient r was evaluated for the 9 significant cytokines. It was observed that the two notable classification contributing cytokines namely IL1-β and IL-10 exhibited moderate positive correlation with a correlation coefficient equal to r = .51. The correlation and significance were showcased in the following graph (**Fig. 15**)

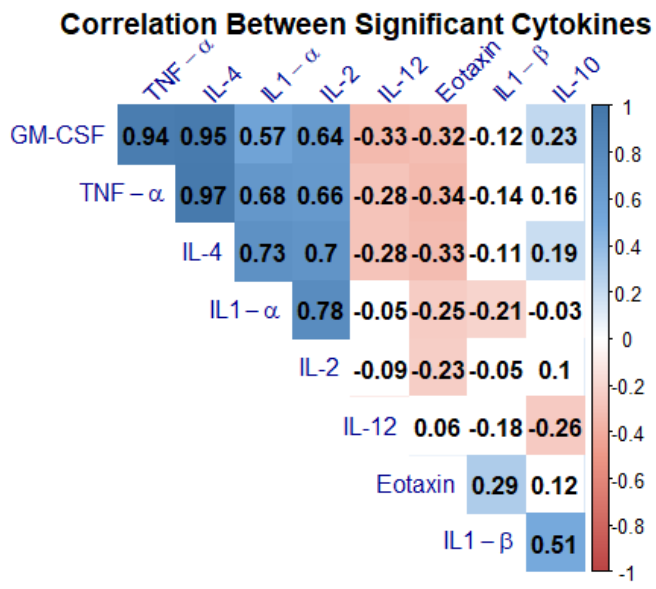


Fig. 15 Cytokine correlation matrix with 9 significant cytokines

9. Classifier 7 Experiment

The final classifier was constructed using k-NN with Canberra distance metric and k=5. The prediction accuracy of .9231 was excellent numerically as well as from the Classifier comparison standpoint. The Classifier 7 upheld the second position after Classifier 6 where prediction accuracy is concerned. The structure of this dataset is multidimensional and is close to zero which conforms to the optimal Canberra distance measure specifications.

Table 15: Classifier 7 Experiment Results for k-NN algorithm with Canberra and k=5

Algorithm	Classification Criterion	Predictor Feature Space	Prediction Accuracy	Sensitivity	Specificity
k-NN	Distance Measure: Canberra And k=5	35 Cytokines	.9231	.8750	1.000

Discussion

The final test results obtained by the classification are synthesized and tabulated in the following table (**Table 16**):

Table 16: Overall Comparison of the testing results for K-NN algorithm and Random Forest.

Classifier Experiment	**Algorithm**	**Classification Criteria**	**Predictor Feature Space**	**Prediction Accuracy**	**Sensitivity**	**Specificity**
Classifier 1	k-NN	Euclidean Distance with k=1	35 Cytokines	0.7059	0.5789	0.7812
Classifier 2	k-NN	Euclidean Distance with k=5	9 Significant Cytokines p-value <.05	0.8293	0.6000	0.9615
Classifier 3	k-NN	Weighted Euclidean Distance with k=8 and inverse kernel	35 Cytokines	0.72	0.5556	0.8125
Classifier 4	k-NN	Weighted Euclidean Distance with k=5 and optimal kernel	9 Significant Cytokines p-value <.05	0.878	0.7333	0.9615
Classifier 5	Random Forest	Distance Trees	35 Cytokines	0.8824	0.7368	0.9688
Classifier 6	Random Forest	Decision Trees	9 Significant Cytokines p-value <.05	1.000	1.000	1.000
Classifier 7	k-NN	Canberra	35 Cytokines	.9231	.8750	1.000

As per the numerical metric performance obtained for cytokine biomarker feature predictors (**Table 16**), the final classification ranged from good to excellent. Classifier 1 experiment (**Table 9,Table 16**) implemented k-NN with Euclidean distance provided an acceptable prediction accuracy of .7059 but was lower than all the other six Classifiers. Classifier 2 experiment using k-NN with Euclidean distance and only significant cytokines obtained a prediction accuracy of .8293 (**Table 10, Table 16**).This showcases an improvement when only nine significant cytokines were included. Classifier 2 performed better than Classifier 1 and 3 but worse than the other three Classifiers. The primary reason for this improvement was the elimination of noise present due to redundant predictor features. The Classifier 3 experiment using weighted k-NN algorithm (**Table 11, Table 16**) evinced an improved prediction accuracy of .72 as compared to the Classifier 1 Experiment, a simple k-NN algorithm

implementation with prediction accuracy of .7059.This improvement was seen because giving precedence to nearest neighbors provided a stronger classification Compared to all classifiers except for Classifier 2, Classifier 3's prediction accuracy exhibited a degradation. The Classifier 3 performed worse than all Classifiers except Classifier 1 because feature space was composed of 35 cytokines, some of which were redundant and did not contribute much toward the final classification. Classifier 4 experiment applied weighted k-NN (**Table 12,Table 16**), and was further enhanced when only significant cytokines were used. It achieved a prediction accuracy of .878 that was better than Classifier 1,2, and 3 but worse than Classifier 5,6, and 7. The Classifier 7 experiment with Random forest (**Table 13,Table 16**) showcased a noteworthy increase in the prediction accuracy, with a value of .8824 compared to k-NN implementations of Classifier 1,2,3, and 4. This was due to the inherent strength of Random Forest that creates multiple decision trees with random feature partition criterion. The degradation of Random forest prediction accuracy due to noisy data is slower compared to k-NN. Classifier 6 experiment (**Table 14,Table 16**) that entailed Random Forest including nine significant cytokines generated a perfect prediction accuracy of 1.00 which surpassed the prediction accuracy of all the other experiments. This was due to the aforementioned strengths of Random forest as well as elimination of noise from the data achieved by dropping redundant predictor features. The final Classifier 7 experiment (**Table 15,Table 16**) which implemented k-NN using the Canberra distance provided exceptionally good prediction accuracy of .9231 which was better than all the other Classifiers except Classifier 7,even though cross validation was not used for it.

Random Forest with significant cytokines (**Classifier 6 Experiment**) overall proved to be the best ML classification technique as it provided a 100% prediction accuracy in conjunction with a 100% of Specificity ,Sensitivity and AUC. To obtain insights pertaining to the contributions of individual cytokines that generated a perfect prediction accuracy, an evaluation was performed that delineated the attributable accountability on a scale of zero to 100. As per the results recorded in Classifier 6 Experiment (**Figure 3**), it was observed that amongst the significant cytokines differentiated by CAD and Control, IL1-β and IL-10 (**Figure 14**) were the most prominent predictor cytokines that can be held accountable for the outstanding evaluation measures. To further explore the existing interrelationships between the significant cytokines that were predominantly responsible for the classification, a correlation matrix (**Figure 15**) was constructed and the significant correlations were identified using the significance threshold value of α-value =.05. It was very interesting to see that the major classification contributors, IL1-β and IL-10, were moderately positively correlated (r=.51, Fig. 15) with each other. IL-10 is considered anti-inflammatory and IL1-β is considered to play an anti-inflammatory role. [5] [20] [21]. Individually, IL-10 exhibited a weak though significantly negative correlation with IL-12 (r=-.26,Fig. 15) and a weak though significantly positive correlation with GM-CSF(r=.23 , Fig. 15). Additionally, IL-4(r=.19 ,Fig. 15)

. IL1-β showed a weak though significantly positive correlation (r=.29,Fig. 15) with Eotaxin and a weak, though significant negative correlation with IL1-α(r=-.21,Fig. 15).

Conclusions

This research uniquely implements the usage 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 diverse k-NN distance measures were compared in terms of efficacy as well as juxtaposed relative to the performance of Random Forest algorithm. Overall, the use of significant cytokines improves the prediction accuracy across all 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 though Random Forest was minimally affected. As compared to prior research studies [6], [9] that used Random Forest with cytokines to differentiate disease groups from control groups, our study exhibited better prediction accuracy(100%) and AUC measures. The improved accuracy is attributable to the preselection of cytokines that were significant across CAD versus Control as well as the use of resampling cross validation.

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 [10] . The Random Forest implementation specifically with significant cytokines provided 100% prediction accuracy outperforming the previous research studies [22]. Finally, our observations implement the Canberra similarity distance metric in the context of the k-NN algorithm, a perspective rarely used in prior research. It also demonstrates a better prediction accuracy than even Random Forest with cross validation. The Canberra distance used in this research study outperformed numerous experimental models for this study in conjunction with previous studies using k-NN with other distance measures [10] [12]. The statistical measures such as correlation coefficient and t tests help identify the complex interactions of the cytokines and narrow the cytokines that are significant in the classification of CAD versus Control.

The main challenge in deciding upon Machine Learning is the trade-off between interpretability and predictive accuracy. Overall, both k-NN and Random Forest gave reasonably good results. Therefore, the algorithm choice depends on contrasting factors such as simple versus complex algorithm, easy to implement versus difficult to implement, and almost perfect accuracy versus reasonable accuracy. Both k-NN and Random Forest balance the trade off with regards to interpretability, but Random Forest provided superior prediction accuracy metrics.

In this age of innovation and all-pervasive machine learning systems, it is important to leverage the abstraction, generalization, optimization, and computational power of the versatile Machine Learning algorithms that can be used across a wide spectrum of domain areas. From the computational standpoint, the distance measures like

Canberra, Cosine, and Correlation can be explored for k-NN, optimized by parameter fine-tuning, and widely proven resampling techniques like Bootstrap and cross-validation.

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

AUC-ROC: Area Under the Curve-Receiver Operating Characteristics

CAD: Coronary Artery Disease

FNR: False Negative Rate

FP: False Positive

k-NN: K Nearest Neighbor

ML: Machine Learning

TN: True Negative

TP: True Positive

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] "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.".

[10] "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.".

[11] "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.".

[12] "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.".

[13] "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.".

[14] "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.".

[15] "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.".

[16] "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.".

[17] "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".

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

[19] "Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot".

[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] "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".

[22] "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.".

521

522 **********