Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

TOWARDS A PERSONALIZED ATHEROSCLEROSIS RISK PREDICTION USING MACHINE LEARNING

*[1]Nentawe Yusuf Gurumdimma, [1]Dimka Betty Toyin, [1]David Enekai Oguche, [2]Sirisena Anil

[1]Department of Computer Science, Faculty of Computing, University of Jos, Jos, Nigeria
[2]Department of Radiology, Jos University Teaching Hospital, Jos, Nigeria

*Corresponding Author Email Address: yusufn@unijos.edu.ng

ABSTRACT
Accurate and early prediction of atherosclerosis and cardiovascular disease (CVD) is essential for effective intervention. While numerous machine learning approaches have been proposed for this task, the majority rely heavily on lab-based (invasive) clinical variables. These lab-dependent methods often involve delayed results and pose accessibility challenges due to their cost and procedural discomfort. In this study, we develop and evaluate a machine learning framework for predicting atherosclerosis risk using both non-laboratory (non-invasive) and laboratory-based clinical indices. We compare the performance of three classification algorithms – Random Forest, Ensemble (Voting) Classifier, and Multilayer Perceptron - across different input configurations. Experimental results demonstrate that the Random Forest classifier achieved an F-Measure of 95%, AUC of more than 98% using only non-lab features, outperforming the use of lab-based features configurations across all models by at least. 5%. These findings highlight the potential of deploying non-invasive, machine learning-based risk assessment tools as point-of-care applications, enabling early prediction of atherosclerosis without the need for laboratory testing.

Keywords: Cardiovascular Disease, Artificial Intelligence, Machine Learning, Prediction, atherosclerosis risk , Non-invasive**.**

INTRODUCTION

Artificial Intelligence has greatly disrupted so many fields(Al Kuwaiti et al., 2023; Ali et al., 2023; Sun et al., 2025) including the healthcare sector. Machine learning is an aspect of artificial intelligence where computers are programmed to learn from data. The development of some of these underlying algorithms relies heavily on computational statistics(Munger et al., 2021). Cardiovascular disease (CVD), including heart attack and stroke, is the leading cause of death worldwide and is usually preceded by accelerated atherosclerosis(Jebari-Benslaiman et al., 2022). Atherosclerotic disease is caused by hardening and narrowing of the arterial walls, which silently and slowly block arteries, putting blood flow at risk due to plaque that is made up of fat, cholesterol, calcium, and other substances that build up or clog the arteries(Björkegren & Lusis, 2022). It begins with damage to the endothelium. Some of the known causes are high blood pressure, smoking, and high cholesterol, among others. When bad cholesterol, or Low-density Lipoprotein (LDL), crosses the damaged endothelium, the cholesterol enters the wall of the artery, which causes the white blood cells to stream in, to digest the LDL. Over the years, cholesterol and cells become plaque in the wall of the artery. Plaque creates a bump on the internal lining of the artery. As atherosclerosis progresses, the artery becomes more and more narrow, which could significantly affect the blood supply to tissues. As a result, not only is the heart at risk, but it can result in heart attacks, strokes, and peripheral vascular disease, which together are called cardiovascular disease (CVD) and cause of other health problems (Jebari-Benslaiman et al., 2022).

The effectiveness of ML in detecting and predicting many diseases, including cancers, infectious diseases, etc., has been established by many researchers (Al Kuwaiti et al., 2023; Ali et al., 2023; Munger et al., 2021; Nikan et al., 2016). Particularly, works have been done in predicting atherosclerosis and CVDs using machine learning; however, these works relied more on lab-based data (Ding et al., 2023; Miranda & Adiarto, 2024; Nikan et al., 2016; Terrada et al., 2020). The challenge with these approaches is that patients must present themselves to the health centres and get tested in the labs to know their risk levels, and the methods rely heavily on lab-based attributes or indices for prediction. This is large due to the belief that lab variables are the most accurate. Additionally, such methods become difficult to implement in low-resource settings for point of care. This is because it will require an implementation that integrates laboratory equipment. This equipment is expensive and not readily available.

In this work, we explore the potential of building a point-of-care method or an approach through which users could know their risk level without having to do laboratory tests. In this approach, we explore the potential of using highly correlated non-laboratory indices for predicting the risk of having atherosclerosis. We compare the results of the laboratory indices with the non-lab indices using the different ML algorithms.

The prediction of cardiovascular diseases (CVD), particularly atherosclerosis, has seen significant advancements through machine learning techniques in recent years. Researchers have explored both invasive and non-invasive approaches to improve early detection and risk assessment. Below is an updated synthesis of key contributions in this domain, incorporating the latest studies.

Machine Learning Approaches for CVD Prediction

Researchers have developed a machine learning-based algorithm to predict coronary artery atherosclerosis using the STULONG and UCI databases (Nikan et al., 2016). Their methodology employed Ridge Expectation Maximization (REM) for missing value imputation, Conditional Likelihood Maximization (CLM) for feature selection, and Extreme Learning Machine (ELM) alongside Support Vector Machine (SVM) for classification, achieving an accuracy of 89.86%. In a comparative study, six data mining tools (Orange, Weka, RapidMiner, Knime, Matlab, and Scikit-learn) and six machine learning techniques (Logistic Regression, SVM, K-Nearest Neighbors, Artificial Neural Network, Naïve Bayes, and Random Forest) were used on a dataset of 303 instances with 13 features. Their results showed that the Artificial Neural Network

(ANN) achieved the highest accuracy (85.86%) when implemented in MATLAB.

(Ayatollahi et al., 2019) compared ANN and SVM for predicting coronary artery disease (CAD) using medical records from three hospitals. The SVM model demonstrated superior performance, with a lower Mean Absolute Percentage Error (112.03), higher Hosmer-Lemeshow test score (16.71), and greater sensitivity (92.23%). The study concluded that SVM provided better goodness-of-fit (74.42) and higher accuracy than ANN. Similarly, (Jindal et al., 2021)applied K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest on a dataset of 304 patients with 14 attributes, mostly lab-based, achieving an accuracy of 88.5% with KNN and Logistic Regression.

Recent Advances in Hybrid and Ensemble Techniques

Recent studies have explored hybrid and ensemble methods to enhance prediction accuracy. For instance, (Singh et al., 2025), (Mehta & Aneja, 2025)proposed a deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to analyze electrocardiogram (ECG) data for CVD prediction, achieving an accuracy of above 97%. Their work highlighted the potential of deep learning in processing complex biomedical signals. Similarly,(Shah et al., 2025) introduced a hybrid model integrating Random Forest and Gradient Boosting Machines (GBM) with a performance (ROC-AUC =0.82).(Latha & Jeeva, 2019)investigated ensemble techniques (Boosting, Bagging, Stacking, and Majority Voting) on the Cleveland heart disease dataset, demonstrating accuracy improvements of up to 7.62%. More recently, multi-modal stacking ensembles were used (Jindal et al., 2021; Shah et al., 2025; Yoon & Kang, 2023). In their work, (Yoon & Kang, 2023) used a combination of ResNet-50 and logistic regression to achieve an AUC of 0.995, an accuracy of 93.97%on 12-lead electrocardiogram (ECG) databases. Their findings emphasized the robustness of ensemble methods in handling heterogeneous medical data.

Non-Invasive Predictors and Wearable Technology

The shift toward non-invasive predictors has gained momentum in recent years. (Sirisena et al., 2022) developed a risk assessment tool for atherosclerotic CVD using six non-invasive indices, demonstrating its efficacy in early screening. (Khozeimeh et al., 2022) utilized a non-invasive approach using cardiac magnetic resonance data and applied Random Forest with CNN features for CAD diagnosis, achieving an accuracy of 0.99. Though their approach performed well, it is used as a precursor to invasive testing.

Another notable contribution by (Bisna et al., 2025)leveraged retinal fundus images and a deep learning framework to predict cardiovascular disease, achieving a good sensitivity. This non-invasive method offered a promising alternative to traditional diagnostic techniques; however, it utilizes images of the retina different from what we use and are difficult to utilize as a point-of-care solution. Similarly, (Kuo et al., 2025) integrated demographic, lifestyle, and genetic data into a federated learning model to predict CVD risk across diverse populations while preserving data privacy, reported high accuracy. The approach focused on data privacy rather than making it a personalized point-of-care solution to CVD. Other similar approaches are reported in (Gill et al., 2023; Kapila & Saleti, 2025). They are largely targeted towards data privacy preservation.

Challenges and Future Directions

Despite these advancements, challenges remain, including data heterogeneity, model interpretability, and the need for large-scale validation. Recent works(Ashika & Hannah Grace, 2025; Talukder et al., 2025)addressed explainability and interpretability by developing a SHAP (Shapley Additive exPlanations)-based explainable AI model for CVD prediction, which provided clinically actionable insights alongside predictions. Future research should focus on integrating multi-modal data (e.g., genomics, proteomics, and imaging) and advancing federated learning frameworks to enhance generalizability and privacy. Another critical approach is allowing individuals to know their risk level early enough to enable early intervention and prevention of deaths. Utilizing non-invasive and easy-to-get data to be used as a point of care is an open area that needs attention. Such data that can point to CVD risk can be used to build risk predictive tools and applications that can be used without needing an expert. This work tries to provide answers to some of the early considerations of this challenge.

The field of CVD prediction has evolved significantly, with machine learning models achieving high accuracy through hybrid and ensemble techniques. Recent studies highlight the potential of non-invasive methods, wearable technology, and explainable AI to improve early detection and patient outcomes. However, further research is needed to address scalability, interpretability, and real-world deployment challenges. Moreso, the non-invasive or the utilization of non-lab variables would facilitate implementation of risk level prediction in low-resource settings. This can totally eliminate the expensive lab processes and can be used by individuals to monitor their risk levels. Our methodology demonstrates how we achieve this.

MATERIALS AND METHODS

Data Collection

The dataset used in this study was obtained from the unpublished Ph.D. thesis of one of the co-authors, which was submitted to the University of Jos [17]. The research received ethical approval from the Institutional Review Board of Jos University Teaching Hospital (JUTH), Nigeria. Data was collected from patients at the JUTH over a period of time, and was investigated by experts for proper labelling.

This comprehensive dataset contains 426 patient records, each characterized by 34 clinical attributes. The variables encompass:

a. Demographic Information: Basic patient characteristics and personal details
b. Non-Laboratory Clinical Measurements: Results from physical examinations and non-invasive tests, recorded either independently or with medical assistance
c. Laboratory Test Results: Biochemical and haematological parameters from diagnostic tests
d. Medical History: Comprehensive records of past diagnoses, treatments, and health conditions

The dataset integrates both objective clinical measurements and subjective patient-reported information, providing a robust foundation for cardiovascular disease risk assessment. The detailed data description is in Table 1.

Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Table 1: The dataset feature description

ATTRIBUTES	DESCRIPTION	ATTRIBUTE TYPE	TYPE
AGE	Length of time a person has lived or existed	Non-Lab	numeric
BODY MASS	Body mass in kilograms	Non-Lab	numeric
HEIGHT	Measurement of a person from head to foot	Non-Lab	Real
WC	The smallest circumference of the natural waist, usually just above the belly button	Non-Lab	Real
AH	Abdominal Height is the distance from the small of the back to the upper abdomen is defined as the thickness of the abdomen at waist level	Non-Lab	numeric
BMI	Body Mass Index is a person's Body mass in kilograms divided by the square of height in meters. A high BMI can indicate high body fatness, and a low BMI can indicate too low body fatness.	Non-Lab	Real
WHtR	Waist-to-Height Ratio(The distribution of body fat). Higher values of the waist-height ratio indicate a higher risk of obesity-related cardiovascular diseases. The waist-height ratio is a good indicator of the risk of heart attack, stroke or death	Non-Lab	Real
WHR	Waist-to-Hip Ratio (compares the size of the waist to the size of the hips in inches) Research shows that people with "apple-shaped" bodies (more weight around the waist) face more health risks than those with "pear-shaped" bodies (more weight around the hips)	Non-Lab	Real
BSI	Body Surface Index is Body mass (kg)/Body surface area (square meter)	Non-Lab	Real
GENDER	Sex (Male/Female)	Non-Lab	Real
RISK	Indicate the Risk level(where 1=Very High Risk, 2=High Risk , 3=Average Risk, 4= Low Risk, 5= Very low Risk)	Non-Lab	Real
SBP	Systolic Blood Pressure. The number on top when measuring blood pressure. This refers to the amount of pressure in the arteries during the contraction of the heart muscle	Non-Lab	numeric
DBP	Diastolic Blood Pressure. The number at the bottom when measuring blood pressure. This is the pressure in the arteries when the heart rests between beats.	Non-Lab	numeric
FBG	Fasting Blood Glucose Test	Lab	Real
TC	Total Cholesterol in the Blood	Lab	Real
TRIG	Level of Triglycerides in the Blood	Lab	Real
RCCLD	Right Common Carotid Luminal Diameter	Lab	Real
RCCIMT	Right Common Carotid Intima-media Thickness	Lab	Real
RCCPSV	Right Common Carotid Peak Systolic velocity	Lab	Real

Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

RCCEDV	Right Common Carotid End-diastolic Volume	Lab	Real
RCCPI	Right Coronary Cusp Pulsatility Index	Lab	Real
PI RISK	Pulsatility Index Risk	Lab	numeric
RCCRI	Right Common Carotid Resistive Index	Lab	Real
OAPSV	Ophthalmic Artery Peak Systolic Velocity	Lab	Real
OAEDV	Ophthalmic Artery End-Diastolic Velocity	Lab	Real
OAPI	Ophthalmic Artery Pulsatility Index	Lab	Real
OARI	Ophthalmic Artery Resistive Index	Lab	Real
SMOKING	Inhale and exhale fumes from burning materials	Non-Lab	numeric
ALCOHOL	A liquid or drink produced by fermentation (containing ethanol)	Non-Lab	numeric
EXERCISE	Physical activity	Non-Lab	numeric
CVD HISTORY	Cardiovascular Disease (Family History)	Non-Lab	numeric
CLASS	Output: the risk level	Output	Nominal

Feature Selection

This is section is aimed at demonstrating the importance of non-lab attributes towards predicting atherosclerosis. The data reduction techniques applied to the dataset were attribute selection using Correlation Based Features Selection by the attribute evaluator with Best First as the search method. This algorithm was chosen because it can measure the linear relationship of one to multiple variables and predict one variable from the other based on its ability to evaluate the worth of a subset of attributes by considering the individual predictive of each feature along with the degree of redundancy between them. Correlation was used for feature selection because it selected relevant (good) variables that are highly correlated with the target class. The selection output 8 attributes from a total of 30 attributes listed in the table below (5 non-lab and 3 lab variables):

```
Attribute Selection on all input data
CFS Subset Evaluator Including locally predictive attributes
Selected attributes: 3,4,8,11,15,18,24: 7
        HEIGHT
        WC
        WHR
        DBP
        RCCLD
        RCCEDV
        OAPI
```

Figure 1: Attributes selected using Correlation-Based Features Selection

The second phase of selection was done based on the method of attribute selection, executing Classifier Attribute Evaluation using Ranker as the search method on the attributes. This technique evaluated the worth of each attribute in the dataset in the context of the output variable (the class). The Ranker was used based on the ability to navigate different combinations of attributes and arrived at a short list of the chosen features. The first 16 attributes ranked were all non-lab indices displayed in the table below. The non-lab attributes were inverted from all the attributes, and the irrelevant attributes were removed using the Preprocess tab window.

```
Ranked attributes:
 0  32 CVD HISTORY
 0  31 EXERCISE
 0  10 GENDER
 0  11 BSIz
 0  12 RISK
 0  13 SBP
 0  14 DBP
 0   9 BSI
 0   8 WHR
 0   7 WHtR
 0   3 HEIGHT
 0   2 BODY MASS
 0   4 WC
 0   6 BMI
 0   5 AH
```

Figure 2: Attributes selected using Classifier Attribute Evaluation

Principal Component Analysis was used for and Ranker as a search method, (see table below) was applied as the third selection, also output non-lab indices as the first group of attributes. The algorithm was used to further validate the selected attributes to help in overcoming data overfitting issues that usually

Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

accompany dimension reduction.

```
Ranked attributes:
 0.8786     1   -0.337AH-0.327WHtR-0.288BMI-0.278WHR-
0.248WC...
 0.7953     2 0.354BSIz+0.31 CLASS=1+0.302RISK =1-0.29BSI-
0.283WC...
 0.7152          3    -0.385PI    RISK-0.365RCCPI-
0.365RCCRI+0.286RCCEDV-0.267BODY
          MASS...
 0.6474          4    0.299BMI+0.281AH+0.272WHtR-
0.249BSIz+0.245DBP...
 0.5902     5 0.51 CLASS=2+0.386RISK =2-0.274RISK =4-
0.268CLASS=4-0.262RISK =1...
 0.5387                    6    -0.373OARI-
0.365OAPI+0.317SBP+0.299OAEDV+0.283DBP...
 0.4921     7  -0.378RCCEDV-0.353OAPSV-0.323CLASS=3-
0.309RCCPSV+0.258RISK =2...
 0.4504   8 -0.45OAEDV-0.347OAPSV-0.292CLASS=5+0.285TC-
0.279RISK =5...
Selected                                   attributes:
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,2
5,26: 26
```

Figure 3: Attributes selected using Principal Components

Machine Learning for Risk Prediction

The Machine Learning algorithms we employed for the prediction are: Random Forest, Multilayer Perceptron, and an Ensemble Approach (a combination of both Perceptron and Random Forest). Random Forest has been proven to be effective in predicting diseases even with minimal data due to its versatility and ability to reduce overfitting. To demonstrate the effectiveness of non-lab and lab attributes in the prediction of the disease, selected these algorithms that have been shown to have performed effectively by other researchers, largely on lab-based attributes. In this work, non-lab variables can also be referred to as non-invasive variables. The Ensemble method utilized in the research is the Voting. It is a method that aggregates the predictions from several models, either by majority rule or by averaging their outputs. We employ hard voting, for which each model in the ensemble makes a prediction, and the final decision is based on the class that gets the most votes across the models. The final output is typically the average of all model predictions.

EXPERIMENT AND RESULTS

Experiment

Experiments were conducted by training the model on the dataset containing 426 records. While performing the experiments, all parameters were set to their default setting, k-fold cross-validation for each algorithm. The performances of the models in this study were evaluated using the metrics: accuracy, precision, recall, and F-measure, which were calculated using the predictive classification table, known as the Confusion Matrix. ROC area was also used to compare the performances of the classifiers.

These experiments were designed to investigate the performance of the classifiers in predicting the risk levels of atherosclerosis and the effect of attribute selection and accuracy.

The experiment involved testing three machine learning models — **Random Forest**, **Voting Classifier**, and **Multilayer Perceptron (MLP)** — using different sets of input features: **Lab variables**, **Non-Lab variables**, and a combination of **both Lab and Non-Lab variables**. The models were evaluated using standard metrics: **Accuracy**, **Precision**, **Recall**, and **F-Measure** (F1 score).

The Results

Random Forest performed exceptionally well across all configurations, with the highest accuracy and F1 score when trained on both Lab and Non-Lab variables. The improvement from using both sets of features is consistent and indicates that the combination provides more predictive power. Non-Lab variables alone slightly outperformed Lab-only variables

Table 2: Results of Experiment with Random Forest

Input Variables	Accuracy	F-Measure	Precision	Recall
Lab only	93.89%	0.939	0.941	0.939
Non-Lab only	94.83%	0.948	0.951	0.948
Both	**95.77%**	**0.958**	**0.958**	**0.958**

The MLP performed the weakest when only a single type of variable (Lab or Non-Lab) was used as can be seen in Table 4.2. Its performance improved significantly with both variable sets, indicating some sensitivity to data richness, but it still fell short of both the Random Forest and the Ensemble (Voting) Classifier on all metrics.

Table 3: Results of Experiment with MLP

Input Variables	Accuracy	F-Measure	Precision	Recall
Lab only	61.50%	0.601	0.596	0.615
Non-Lab only	61.73%	0.604	0.607	0.617
Both	83.56%	0.832	0.835	0.836

The Ensemble (Voting) Classifier showed moderate performance when using only Lab or Non-Lab variables individually. However, performance improved significantly when both sets were combined — showing that this ensemble approach benefits greatly from richer feature input. Still, even at its best, it did not outperform the Random Forest (see Table 4).

Table 4: Results of Experiment with the Ensemble (Voting) Method

Input Variables	Accuracy	F-Measure	Precision	Recall
Lab only	73.00%	0.722	0.731	0.730
Non-Lab only	77.93%	0.773	0.789	0.779
Both	90.14%	0.900	0.904	0.901

Over all, Random Forest shows best performance across all input data. The combination of invasive and non-invasive variables achieved a better performance. However, detail result shows that the use of only the non-lab or non-invasive variables performs better than the use of lab variables across all models.

Table 5: Comparison between all models used.

Model	Best Accuracy	Best F-Measure	Best Input Type
Random Forest	**95.77%**	**0.958**	Both Lab & Non-

Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Model	Best Accuracy	Best F-Measure	Best Input Type
			Lab
Voting Classifier	90.14%	0.900	Both Lab & Non-Lab
Multilayer Perceptron	83.56%	0.832	Both Lab & Non-Lab

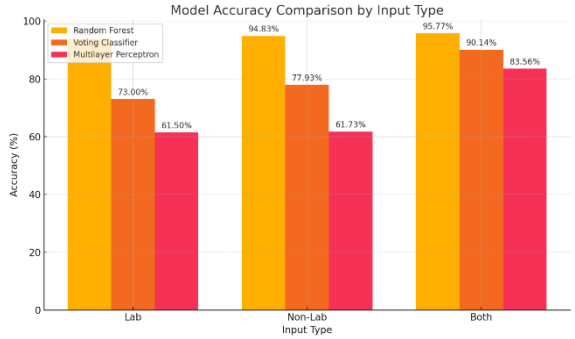

Figure 4.: Model Accuracy comparison by Variable input type

The use of non-Lab variable input types for predicting atherosclerosis has shown great results at par with lab variables, as seen in Figure 4. This is a significant result going forward, as systems can be built based purely on non-lab variables that can enable individuals to know their risk level without going through invasive methods.

The ROC & AUC

The ROC for the three models is shown in Figures 4.2a to 4.4c. The graphs show the true positive rates (Sensitivity) and true negative rates for RandomForest, MLP, and Ensemble (Vote) models. The graphs show the use of lab, non-lab variables, and a combination of both. The AUC for these models and data demonstrates good prediction performance. RandomForest with both lab and non-lab variables has an AUC=0.996; non-lab variables have an ROC=0.994. this shows that there is no significant performance from using both lab and non-lab variables. This can be seen on other models used as well.

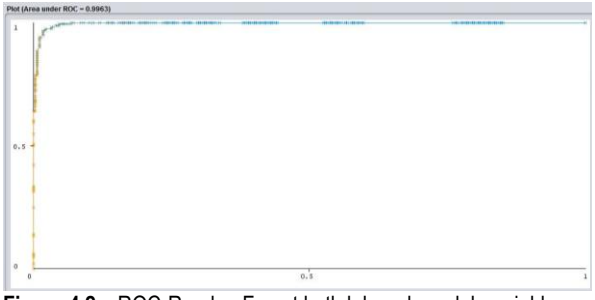

Figure 4.2a: ROC-RandomForest both lab and non lab variable

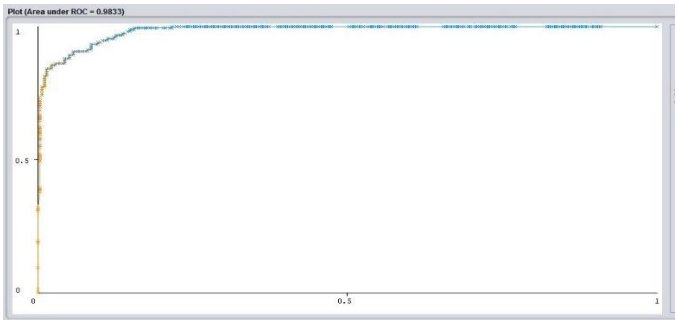

Figure 4.2b: ROC- RandomForest Lab only

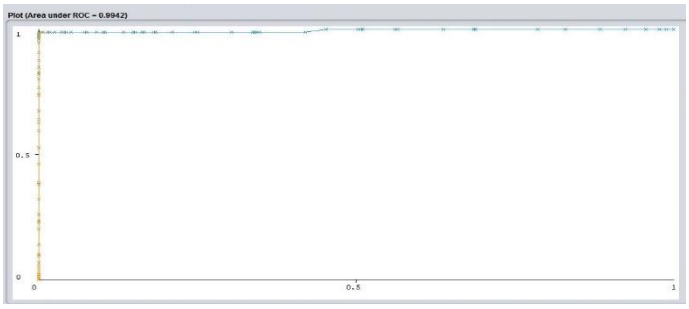

Figure 4.2c: ROC- Random Forest (Non-lab variables only_)

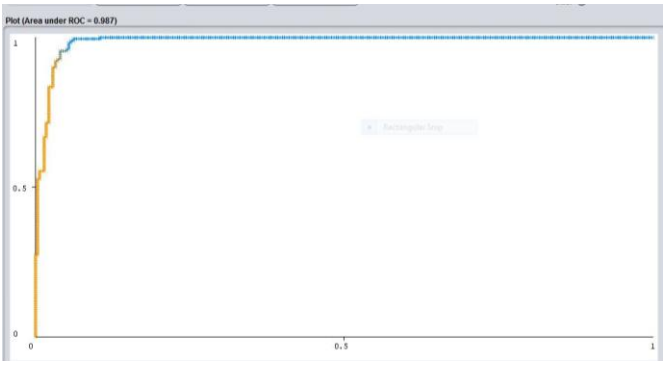

Figure 4.3a: ROC- MLP (both Lab and Non Lab variables)

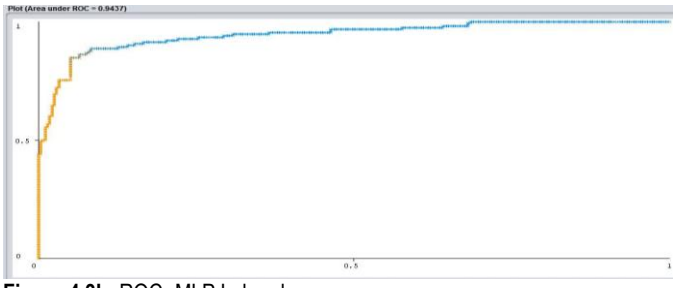

Figure 4.3b: ROC- MLP Lab only

Science World Journal Vol. 20(No 4) 2025 https://dx.doi.org/10.4314/swj.v20i4.36
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

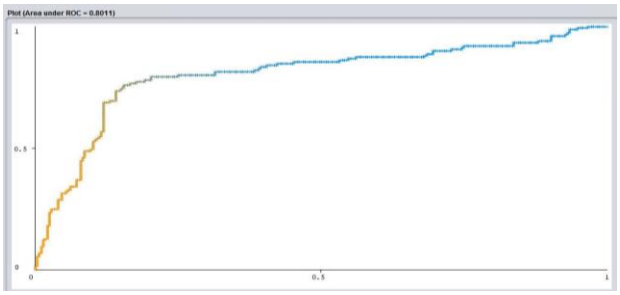

Figure 4.3c: ROC-MLP Lab only

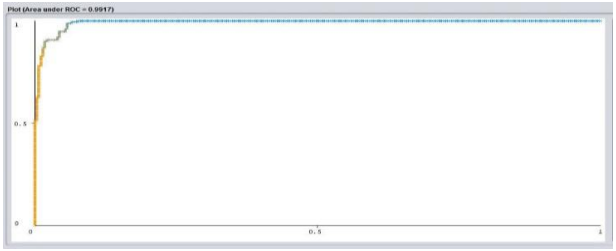

Figure 4.4a: ROC-Ensemble (Vote) lab and non-lab variables combined

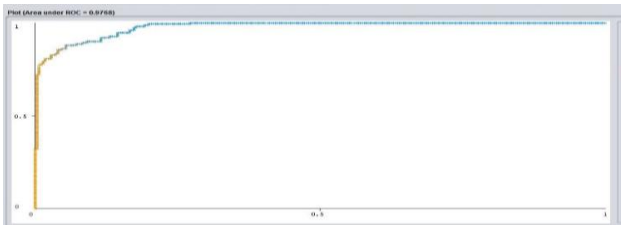

Figure 4.4b: ROC- Ensemble (Vote) Lab variables only

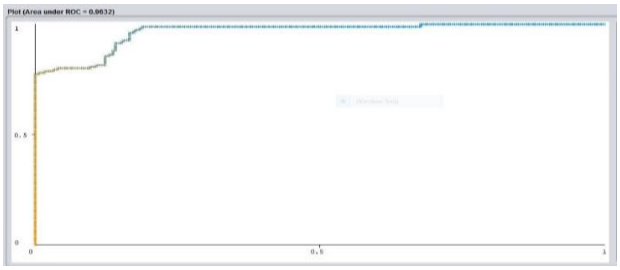

Figure 4.4c: ROC- Ensemble (Vote) Non lab Variables

DISCUSSION AND CONCLUSION

In this paper, we explored how well machine learning models can predict the risk of atherosclerosis using different types of input variables – those obtained through lab tests and those that do not require any invasive procedures. Interestingly, the findings show that non-lab (non-invasive) variables are not only highly informative but, in some cases, even outperform lab-based features in predicting risk levels.

Among the models tested, Random Forest stood out consistently, achieving the best results across all metrics and feature sets. When both lab and non-lab variables were combined, the model reached its highest accuracy of 95.77%, and an AUC=0.996, suggesting that integrating diverse sources of data yields the most reliable predictions. However, it's worth emphasizing that even when limited to non-lab variables alone, Random Forest still performed exceptionally well, achieving 94.83% accuracy, AUC=0.994. This result is significant – it shows that a non-lab variable can be a better data source for predicting Arteriosclerosis without needing an invasive approach.

This pattern was not unique to Random Forest. The Ensemble (Voting) classifier also performed better with non-lab variables (Accuracy =90.14%, AUC=0.99), than with lab-only features, and even the Multilayer Perceptron (MLP), though generally less effective, showed improvement when both variable types were combined. Notably, across all models, non-lab features consistently led to higher scores than lab-only inputs.

These findings carry meaningful implications. Non-lab variables—such as age, blood pressure, smoking status, or BMI-are easy to collect and don't require medical equipment or invasive tests. In a developing world where access to healthcare can be uneven and cost is a major barrier, the ability to accurately assess cardiovascular risk using just these non-invasive inputs is a major advantage. It opens up possibilities for wider, more equitable access to preventive care, especially in rural or underserved regions.

Perhaps most striking is the fact that the performance difference between using only the non-lab variables versus the full set of features was relatively small. This suggests that while lab data can add value, they may not be strictly necessary for building an effective risk prediction tool. With appropriate models, we can develop systems that are both accessible and accurate—enabling people to assess their cardiovascular risk from the comfort of their own homes, perhaps even using wearable devices or mobile health apps.

In conclusion, this study supports the idea that non-invasive, easily measurable data can play a central role in the future of cardiovascular risk prediction. While there's always room for improving model robustness and generalizability, the results here point to a promising direction: intelligent, user-friendly tools that empower early intervention without the need for lab-based diagnostics.

REFERENCES

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D., & Al-Muhanna, F. A. (2023). A Review of the Role of Artificial Intelligence in Healthcare. *Journal of Personalized Medicine*, *13*(6), 951. https://doi.org/10.3390/jpm13060951

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M. A. A., & Dwivedi, Y. K. (2023). A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*, *8*(1), 100333. https://doi.org/10.1016/j.jik.2023.100333

Ashika, T., & Hannah Grace, G. (2025). Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: An optimized RST-ML approach. *Frontiers in Digital Health*, *7*, 1609308. https://doi.org/10.3389/fdgth.2025.1609308

Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: A comparison between two data mining algorithms. *BMC Public Health*, *19*(1), 448. https://doi.org/10.1186/s12889-019-6721-5

Bisna, N. D., Sona, P., & James, A. (2025). Retinal Image Analysis for Heart Disease Risk Prediction: A Deep Learning

Approach. *IEEE Access*, *13*, 76388–76399. https://doi.org/10.1109/ACCESS.2025.3562433

Björkegren, J. L. M., & Lusis, A. J. (2022). Atherosclerosis: Recent developments. *Cell*, *185*(10), 1630–1645. https://doi.org/10.1016/j.cell.2022.04.004

Ding, J., Luo, Y., Shi, H., Chen, R., Luo, S., Yang, X., Xiao, Z., Liang, B., Yan, Q., Xu, J., & Ji, L. (2023). Machine learning for the prediction of atherosclerotic cardiovascular disease during 3-year follow up in Chinese type 2 diabetes mellitus patients. *Journal of Diabetes Investigation*, *14*(11), 1289–1302. https://doi.org/10.1111/jdi.14069

Gill, S. K., Karwath, A., Uh, H.-W., Cardoso, V. R., Gu, Z., Barsky, A., Slater, L., Acharjee, A., Duan, J., Dall'Olio, L., El Bouhaddani, S., Chernbumroong, S., Stanbury, M., Haynes, S., Asselbergs, F. W., Grobbee, D. E., Eijkemans, M. J. C., Gkoutos, G. V., Kotecha, D., … Center, J. (2023). Artificial intelligence to enhance clinical value across the spectrum of cardiovascular healthcare. *European Heart Journal*, *44*(9), 713–725. https://doi.org/10.1093/eurheartj/ehac758

Jebari-Benslaiman, S., Galicia-García, U., Larrea-Sebal, A., Olaetxea, J. R., Alloza, I., Vandenbroeck, K., Benito-Vicente, A., & Martín, C. (2022). Pathophysiology of Atherosclerosis. *International Journal of Molecular Sciences*, *23*(6), 3346. https://doi.org/10.3390/ijms23063346

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012072. https://doi.org/10.1088/1757-899X/1022/1/012072

Kapila, R., & Saleti, S. (2025). Federated learning-based disease prediction: A fusion approach with feature selection and extraction. *Biomedical Signal Processing and Control*, *100*, 106961. https://doi.org/10.1016/j.bspc.2024.106961

Khozeimeh, F., Sharifrazi, D., Izadi, N. H., Joloudari, J. H., Shoeibi, A., Alizadehsani, R., Tartibi, M., Hussain, S., Sani, Z. A., Khodatars, M., Sadeghi, D., Khosravi, A., Nahavandi, S., Tan, R.-S., Acharya, U. R., & Islam, S. M. S. (2022). RF-CNN-F: Random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance. *Scientific Reports*, *12*(1), 11178. https://doi.org/10.1038/s41598-022-15374-5

Kuo, T.-T., Gabriel, R. A., Koola, J., Schooley, R. T., & Ohno-Machado, L. (2025). Distributed cross-learning for equitable federated models—Privacy-preserving prediction on data from five California hospitals. *Nature Communications*, *16*(1), 1371. https://doi.org/10.1038/s41467-025-56510-9

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Mehta, S., & Aneja, A. (2025). Enhanced ECG Signal Analysis Using Hybrid CNN-LSTM and Augmentation Techniques. *2025 International Conference on Automation and Computation (AUTOCOM)*, 1657–1661. https://doi.org/10.1109/AUTOCOM64127.2025.10956808

Miranda, E., & Adiarto, S. (2024). Enhancing automatic early arteriosclerosis prediction: An explainable machine learning evidence. *Clinical eHealth*, *7*, 153–163. https://doi.org/10.1016/j.ceh.2024.12.003

Munger, E., Hickey, J. W., Dey, A. K., Jafri, M. S., Kinser, J. M., & Mehta, N. N. (2021). Application of machine learning in understanding atherosclerosis: Emerging insights. *APL Bioengineering*, *5*(1), 011505. https://doi.org/10.1063/5.0028986

Nikan, S., Gwadry-Sridhar, F., & Bauer, M. (2016). Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis. *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 34–39. https://doi.org/10.1109/CSCI.2016.0014

Shah, P., Shukla, M., Dholakia, N. H., & Gupta, H. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. *Scientific Reports*, *15*(1), 17927. https://doi.org/10.1038/s41598-025-01650-7

Singh, S., Singh, A., Singh, S., & Khurana, R. (2025). CVLSTMLW-CNN: A IoT-Enabled Hybrid CNN Model for Heart Disease Prediction. In S. N. Mohanty, S. Satpathy, X. Cheng, & S. K. Pani (Eds.), *Explainable IoT Applications: A Demystification* (Vol. 21, pp. 349–358). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74885-1_23

Sirisena, A. U. I., Gurumdimma, N. Y., Oguche, D. E., & Okeahialam, B. N. (2022). Development of a Jos cardiovascular disease risk app to improve screening for atherosclerotic cardiovascular diseases. *Health Technology*, *6*, 1–1. https://doi.org/10.21037/ht-21-22

Sun, H., Cai, Z., Zhou, H., & Tan, C.-W. (2025). *The Role of AI Technologies in Healthcare Routines: A Literature Review*.

Talukder, Md. A., Talaat, A. S., & Kazi, M. (2025). HXAI-ML: A hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection. *Results in Engineering*, *25*, 104370. https://doi.org/10.1016/j.rineng.2025.104370

Terrada, O., Cherradi, B., Raihani, A., & Bouattane, O. (2020). Atherosclerosis disease prediction using Supervised Machine Learning Techniques. *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 1–5. https://doi.org/10.1109/IRASET48871.2020.9092082

Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. *Health and Technology*, *10*(5), 1137–1144. https://doi.org/10.1007/s12553-020-00438-1

Yoon, T., & Kang, D. (2023). Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases. *Journal of Personalized Medicine*, *13*(2), 373. https://doi.org/10.3390/jpm13020373