

The Impact Of Comorbid Conditions On The Progression Of Peripheral Artery Disease And Associated Cardiovascular Risks: A Data-Driven Approach

Nazia Sultana^{1*}, Dr Kumar P K²

^{1*}Research scholar, Department of CSE, Visvesvaraya Technological University, Postgraduate Centre, Mysuru

²Assistant Professor, Department of CSE, Visvesvaraya Technological University, Postgraduate Centre, Mysuru
naziaimedoh@gmail.com, pandralli@gmail.com,

*Corresponding Author: Nazia Sultana

*Email: naziaimedoh@gmail.com

ORCID: 0009-0008-5104-6793

Abstract:

Among the numerous circulatory conditions we assist patients with is Peripheral Artery Disease (PAD), which is highly comorbid with diabetes, anxiety, depression, and smoking. As such, these conditions can lead to the progression of PAD from the legs to the kidneys, resulting in Chronic Kidney Disease (CKD) and even the heart, resulting in heart attacks, as well as, in extreme cases, death from delayed diagnosis and treatment. Data from the National Health and Nutrition Examination Survey (NHANES), August 2021 – August 2023, is used in this study to develop predictive models of early detection of PAD and related cardiovascular risks.

Keywords: Peripheral Artery Disease, Chronic Kidney Disease, Cardiovascular Risk, Predictive Model, Comorbidity.

Introduction:

Peripheral Artery Disease (PAD)[1-2] is not only a common and often overlooked vascular condition, but principally, it is one in which the blood flow to the limbs, especially the legs, is gradually diminished by a form of atherosclerosis that causes the narrowing of the arteries. The same has symptoms that include mild (foot pain, leg cramps, muscle fatigue, etc.) to more serious issues, unless treated, where ulcers, gangrene, or amputation of a limb become possible due to a drop in blood flow. But physicians have started to realize that PAD isn't merely a localized scourge—it's sort of a systemic disease that, in many cases, can be associated with Chronic Kidney Disease (CKD) and heart attacks and strokes in particular.

Although there are many possible contributing factors, such as diabetes, high blood pressure, smoking, and mental health disorders such as depression and anxiety, that have been shown to speed up the progression of PAD, aside from diabetes, which accelerates the disease by promoting inflammation, oxidative stress, and endothelial dysfunction, all of which accelerate the development of arterial plaque, other factors make PAD a serious concern. However, PAD symptoms tend to be more difficult for people with chronic diabetes to detect early because chronic high blood sugar levels in diabetic patients further increase the risk of nerve damage. Like a disease, mental health conditions like depression and anxiety may add to complications in disease management by not following medical advice in the right way, adopting an unhealthy lifestyle, and not following up at the right time. The emphasis and the reduction of motivation can elevate the factors leading to the lack of obligatory alterations of lifestyle, such as a proper diet and exercise, which is significant in the management of PAD. The use of tobacco that continuously harms the arteries through direct injury, facilitates inflammation, and exacerbates the

stiffness of blood vessels is one of the most significant risk factors. The combination of these conditions leads to the acceleration of the development of PAD and the high likelihood of the latter acquiring critical complications, deteriorating the overall health of a patient, and decreasing their life expectancy.

In order to contribute to the enhancement of early diagnosis and individual treatment plans, it is highly essential to have a grasp of the blanket of complicated interrelations between PAD and the interconnected disorders. Healthcare analytics based on data have now been advanced to allow exploration of these relationships to a greater extent. The present study examines a complex relationship between PAD, comorbidity conditions [3], and mental health [4] that has been reported in data obtained on the National Health and Nutrition Examination Survey (NHANES)[6] between August 2021 and 2023. This is a very rich data set (NHANES) provided which provides demographic, health, and mental health data. The paper applies machine learning technologies and statistical modeling to detect high-risk individuals and identify decisive predictors of PAD development in people. The aim is to make a contribution to early intervention, preventive measures, and individual healthcare to enhance patient outcomes.

In addition, this study will approximate the predictive capacity of inflammatory biomarkers [5] such as C- Reactive Protein (CRP), Homocysteine concentrations, and lipid profiles in determining the PAD risk among patients with numerous health issues. A combination of these clinical assessments with lifestyle and patient history might contribute to the effectiveness of diagnostic measures to aid in the treatment effectiveness by using these biomarkers. The goal of this study is to have a holistic approach in order to help enhance the screening tools and interventions. In the end, it emphasizes the immediate necessity to adopt a multidisciplinary approach provided by cardiologists, endocrinologists, mental health professionals, and public health policymakers in the fight against the increasing cost of PAD and its associated cardiovascular risks.

Literature Review

PAD is evolving from its historical understanding as a simple disease involving narrowing arterial blood flow to the lower limbs due to atherosclerosis and inflammation of the lower limb arterial walls. Still, it is a projection of a multisystem atherosclerotic disease complicated by various metabolic-inflammatory, behavioral, and psychosocial factors. It usually starts with a simple intermittent claudication or very mild pain in the calf or leg, but it can further develop into critical limb ischemia, nonhealing ulcers, and amputations. Beyond the limb, PAD is associated with an increased risk of developing CKD, myocardial infarction, and stroke. Given the intricate relationship between the set of biomarkers, it is necessary to provide a qualitative meta-analysis of existing work, based on large-scale epidemiological surveys, biomarker studies, behavioral research, and new machine learning methods to contextualize the current study that uses a data-driven approach. Epidemiological works built on the analysis of the National Health and Nutrition Examination Survey (NHANES) results proved that PAD increases progressively with age and involves almost 15% of patients older than 70 years of age; more than 95% of patients with PAD have at least one conventional cardiovascular risk factor, including smoking, hypertension, diabetes, and hypercholesterolemia [1]. These results highlighted that smoking increases the risk by over 4 times, diabetes by 2.86 times, and hypertension by 1.9 times for the occurrence of PAD. Namely, the three inflammatory markers, C-reactive protein (CRP) and fibrinogen, were independent predictors that highlighted the mechanism of peripheral atherogenesis [2]. Diabetes mellitus, exemplified by type II diabetes, is known to progress PAD through endothelial dysfunction resulting from chronic hyperglycemia and oxidative stress, and the accumulation of advanced glycation end products. Apart from promoting the development of plaques, diabetes camouflages

early signs of PAD through peripheral neuropathy, making the disease progression worse, and ultimately affecting the feet and resulting in ulcers and amputations [3]. At the same time, hypertension tends to cause shear forces in the vessels, while dyslipidemia, which is characterized by high levels of LDL and low levels of HDL, is also likely to lead to plaque formation. Another condition that worsens these metabolic disturbances is obesity and a sedentary lifestyle; hence, changing lifestyle significantly contributes to the prevention of PAD. Psychological issues have been identified as both cause and effect of PAD in today's society. Additionally, depression and anxiety affect treatment regime compliance and lifestyle measures, and they increase neuroendocrine stress, hypothalamic–pituitary–adrenal axis activity, and the level of pro-inflammatory cytokines that can enhance the pathogenic processes of vascular damage [4]. Specialty clinic patient surveys found a range of PAD patients' prevalence of depression between 40 and 46 percent. In contrast, diabetic patients with PAD are twice as likely to die if they have comorbid depression [5]. Since it is a bidirectional relationship, psychological screening must become a part of the vascular care pathways. Compared to traditional lipids, biomarkers of inflammation have been found to have prognostic significance at the molecular level. Ridker et al. found that CRP had a better predictive value than other biomarkers such as fibrinogen, homocysteine, and lipoprotein (a) for developing incident PAD. In contrast, other studies attributed the increase in severity of PAD to leukocyte count and interleukin-6. These markers not only indicate the inflammation of the vessel wall but also incrementally re-risk-stratify when layered on clinical factors. While tools from statistics help in depicting relationships, machine learning has the potential for an early and patient-specific intervention. Another study revealed that the deep learning models based on endocrine and cardiology clinic data have an accuracy and recall of >97% in detecting cardiovascular risks for type 2 diabetes [6]. Previous studies based on the hemodynamic features extracted from NHANES have reported F1 scores of higher than 0.9 using random forest models for PAD classification [7]. These models can include many variables, such as demography, lab results, mental health scores, and lifestyle indices, and create risk scores for screenings and interventions. Nevertheless, gaps remain. There are no published studies that simultaneously model metabolic, inflammatory, behavioral, and psychological factors in a predictive fashion, and the vast majority of machine learning approaches were not validated in other large studies, including independent population cohorts. However, further studies are needed to determine the level of biomarker increases that would constitute clinically relevant additional risk; the effect of combined mental health interventions on the progression of PAD has not been seen conclusively in clinical trials. The present study aims to fill these gaps by using the data collected within the National Health and Nutrition Examination Survey 2021–2023 to design and compare statistical and machine learning models that may help diagnose early PAD and stratify CVD risk based on clinical characteristics, inflammatory markers, lifestyle practices, and mental health scores.

Methodology:

1) Data Collection and Processing:

To detect PAD early and develop a personalized treatment strategy, we need to understand how PAD is connected to each of its conditions. Healthcare analytics based on data have now been advanced to allow exploration of these relationships to a greater extent. Data used to establish the interrelations between PAD, comorbid conditions, and mental health are based on the data of the National Health and Nutrition Examination Survey (NHANES) of August 2021–2023, all available data of the NHANES. This rich dataset (from NHANES) gives demographic information, health measurements, and mental health assessments.

The paper applies machine learning technologies and statistical modeling to detect high-risk individuals and identify decisive predictors of PAD development in people. The aim is to make a contribution to early intervention, preventive measures, and individual healthcare to enhance patient outcomes.

Moreover, the study will attempt to approximate the predictive quality of inflammatory biomarkers such as C- Reactive Protein (CRP), Homocysteine concentrations, and lipid results in the determination of the risks of PAD in individuals with several health issues. The biomarkers can enhance the diagnostic accuracy, clinical, and patient history, which are lifestyle factors. This research seeks to offer an in-depth solution to improve the formulation of better screening and intervention instruments. On the whole, it posits a multidisciplinary solution that involves cardiologists, endocrinologists, mental health practitioners, and policymakers in the field of public health to fight the increasing burden of PAD and all the associated risks of cardiovascular diseases.

2) Data Preparation

Therefore, several preprocessing was done to ensure that the findings were precise and sound.

- **Dataset Integration:** To integrate all the data points into one, a unified identifier, SEQN, of every participant was formed.

Missing Data Handling: There was too much missing data on these variables (more than 60 percent). The rest of the missing values were imputed, but these were not included in the analysis or imputed depending on their relevance to the analysis.

- **Additional Analysis of Health Conditions of Concern** — It also examined the indicators of such health conditions as diabetes, depression, anxiety, smoking, and Peripheral Artery Disease (PAD).

3) Creating Composite Scores

To achieve more analysis, composite scores were made.

- To determine the severity of depression and anxiety through PHQ-9 responses (DPQ010 to DPQ090) and the severity of anxiety through FNQ140 and FNQ510, the calculation of Depression and Anxiety Scores (DPQ010-DPQ090) was done, respectively.
- They developed a Lifestyle Risk Index that adds up the cumulative number of harm lifestyle factors (including smoking and drinking) that are inflicting on the PAD progression.

4) Statistical and Machine Learning Analysis

Patterns and predictive insights were discovered using a combination of statistic tools and machine learning models.

The dataset's distribution of comorbid conditions can be analyzed through descriptive statistics.

Pearson correlation coefficients were used to determine significant relationships between PAD and other health conditions in the correlation analysis.

Machine learning models: For CKD prediction, logistic regression and random forest models were used, and for CVA, logistic regression, random forest, and K-nearest neighbours models were used. The feature selection was based on each variable's predictive importance.

Reliable Predictions: To evaluate the performance of these models, accuracy, precision, recall, F1-score, and ROC-AUC were used to ensure reliable predictions.

5) Data Visualization

Various visual tools were used to make complex relationships easier to interpret.

- Heatmaps and Correlation Matrices: Displayed the strength of relationships between different health variables.

Illustrations of how PAD progresses through CKD and cardiovascular diseases were more transparently illustrated regarding risk pathways.

Data set used

Index	RIDAGEYR	RIAGENDR_2.0	Mean_Systolic_BP	Mean_Diastolic_BP	LBXTC	LBDHDD	LBXGLU	LBXGH	LBXHSCR	BMI	Total_Physical_Activity	Depression_Score	Anxiety_Score
0	43.0	false	132.00000000000000	90.0	264.0	45.0	113.0	5.0	1.78	27.0	80.0	0.0	0.0
1	60.0	false	117.0	70.00000000000000	214.0	80.0	99.0	5.0	2.03	33.5	90.0	1.0	0.0
2	44.0	true	108.0	75.33333333333333	197.0	49.0	150.0	8.2	5.02	26.7	0.0	2.0	0.0
3	5.0	true	NaN	NaN	NaN	NaN	NaN	NaN	NaN	21.0	0.0	0.0	0.0
4	3.0	false	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
5	3.0	true	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
6	43.0	false	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
7	65.0	true	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	150.0	0.0	0.0
8	24.0	false	115.0	73.00000000000000	163.0	40.0	101.0	5.1	1.05	30.2	80.0	1.0	0.0
9	68.0	true	141.33333333333334	76.0	203.0	42.0	NaN	5.0	3.96	42.6	0.0	4.857844812349025e-78	0.0
10	27.0	true	129.0	90.00000000000000	NaN	NaN	NaN	4.0	NaN	43.7	0.0	0.0	0.0
11	58.0	false	103.0	76.33333333333333	NaN	NaN	NaN	5.5	NaN	28.0	75.0	4.857844812349025e-78	0.0
12	31.0	true	115.33333333333333	74.00000000000000	156.0	38.0	NaN	5.0	11.2	48.0	0.0	0.0	0.0
13	33.0	true	106.0	72.00000000000000	157.0	41.0	NaN	5.4	3.55	38.8	30.0	24.0	0.0
14	54.0	true	108.33333333333334	71.33333333333333	191.0	37.0	NaN	5.0	8.40	43.0	0.0	1.0	0.0
15	38.0	false	130.0	83.0	179.0	54.0	NaN	5.5	4.30	48.1	100.0	2.0	0.0
16	51.0	false	105.00000000000000	65.0	183.0	48.0	88.0	4.8	0.82	24.4	75.0	4.857844812349025e-78	0.0
17	33.0	true	110.33333333333333	81.0	NaN	NaN	100.0	5.0	11.83	54.6	0.0	54.0	0.0
18	56.0	true	123.33333333333333	80.00000000000000	214.0	88.0	134.0	5.0	0.94	27.0	0.0	7.0	0.0
19	67.0	true	145.00000000000000	91.33333333333333	NaN	NaN	138.0	8.2	1.41	46.7	0.0	8.0	0.0
20	47.0	false	112.33333333333333	70.0	188.0	41.0	107.0	5.8	5.20	26.7	180.0	4.857844812349025e-78	0.0
21	68.0	false	120.0	70.33333333333333	193.0	82.0	NaN	5.7	8.26	22.7	150.0	1.0	0.0
22	36.0	false	87.33333333333333	40.00000000000000	173.0	42.0	NaN	5.5	4.00	26.0	0.0	2.0	0.0
23	18.0	false	NaN	NaN	NaN	NaN	NaN	NaN	NaN	28.0	0.0	7.0	0.0
24	26.0	true	NaN	NaN	135.0	54.0	73.0	5.2	0.45	NaN	120.0	0.0	0.0
25	12.0	true	100.0	65.0	174.0	47.0	NaN	4.8	8.07	26.7	0.0	0.0	0.0
26	73.0	true	134.33333333333334	75.00000000000000	179.0	33.0	100.0	5.7	13.7	25.7	80.0	4.0	0.0
27	3.0	true	NaN	NaN	NaN	NaN	NaN	NaN	NaN	14.5	0.0	0.0	0.0
28	26.0	true	105.33333333333333	67.00000000000000	141.0	37.0	NaN	5.0	1.88	25.1	120.0	2.0	0.0
29	23.0	true	142.00000000000000	78.33333333333333	231.0	81.0	NaN	5.1	0.24	26.1	90.0	4.857844812349025e-78	0.0
30	80.0	false	132.00000000000000	60.33333333333333	169.0	82.0	99.0	4.8	0.95	21.4	90.0	22.0	0.0
31	68.0	false	119.0	73.00000000000000	163.0	80.0	NaN	5.0	3.94	26.5	120.0	4.857844812349025e-78	0.0
32	77.0	true	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
33	4.0	false	NaN	NaN	NaN	NaN	NaN	NaN	NaN	16.7	0.0	0.0	0.0
34	7.0	false	NaN	NaN	NaN	NaN	NaN	NaN	NaN	21.8	0.0	0.0	0.0
35	67.0	true	120.0	72.00000000000000	217.0	80.0	86.0	5.8	93.86	36.1	0.0	50.0	0.0
36	10.0	false	112.0	71.33333333333333	NaN	NaN	NaN	5.2	NaN	51.9	0.0	0.0	0.0
37	61.0	true	113.00000000000000	62.00000000000000	157.0	43.0	NaN	13.3	18.05	41.3	0.0	52.0	0.0

Figure 1.1: Dataset used

Key observations:

Columns and Attributes:

RIDAGEYR: Age of individuals in the dataset.

RIAGENDR_2.0: Gender information (true/false).

Blood pressure values with some missing (NaN) entries.

Total Cholesterol & HDL: LBXTC / LBDHDD

Glucose level is important for diabetes risk assessment (LBXGLU).

LBXGH / LBXHSCR: Some other blood health indicators.

BMI (Body Mass Index): One of the determinants of obesity and metabolic health.

Total_Physical_Activity: The activity levels can be used to measure lifestyle behavior.

Depression_Score / Anxiety_Score: Psychological well-being measures.

Data Gaps and Missing Values:

There are also some (NaN) values that are missing some of the entries, especially blood pressure and cholesterol, physical activity, and mental health variability in the measured scores.

Potential Uses:

The dataset may be used to forecast such health conditions as Peripheral Artery Disease (PAD), heart disease, or diabetes.

It is used in correlation analysis between the mental well-being (Depression_Score, Anxiety_Score) and the physical well-being indicators.

Summary:

It presents a dataset containing multiple health-related parameters collected at NHANES (National Health and Nutrition Examination Survey). The dataset is composed of a number of columns that contain demographic, clinical measurement, laboratory test, and mental health scores. The age, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, body mass index (BMI), physical activity, and depressive and anxiety scale scores are the most significant areas of the dataset. Many entries will have missing values (NaN), indicative of the fact that not all the subjects had a full account of all the parameters. Such missing values are common in massive health surveys.

The information provided was of great insight into health conditions, lifestyle practices, and potential risk factors of chronic disease. The causes of diseases like diabetes, cardiovascular complications, and obesity have been known to be a result of high cholesterol, high blood glucose, high BMI, and low exercise.

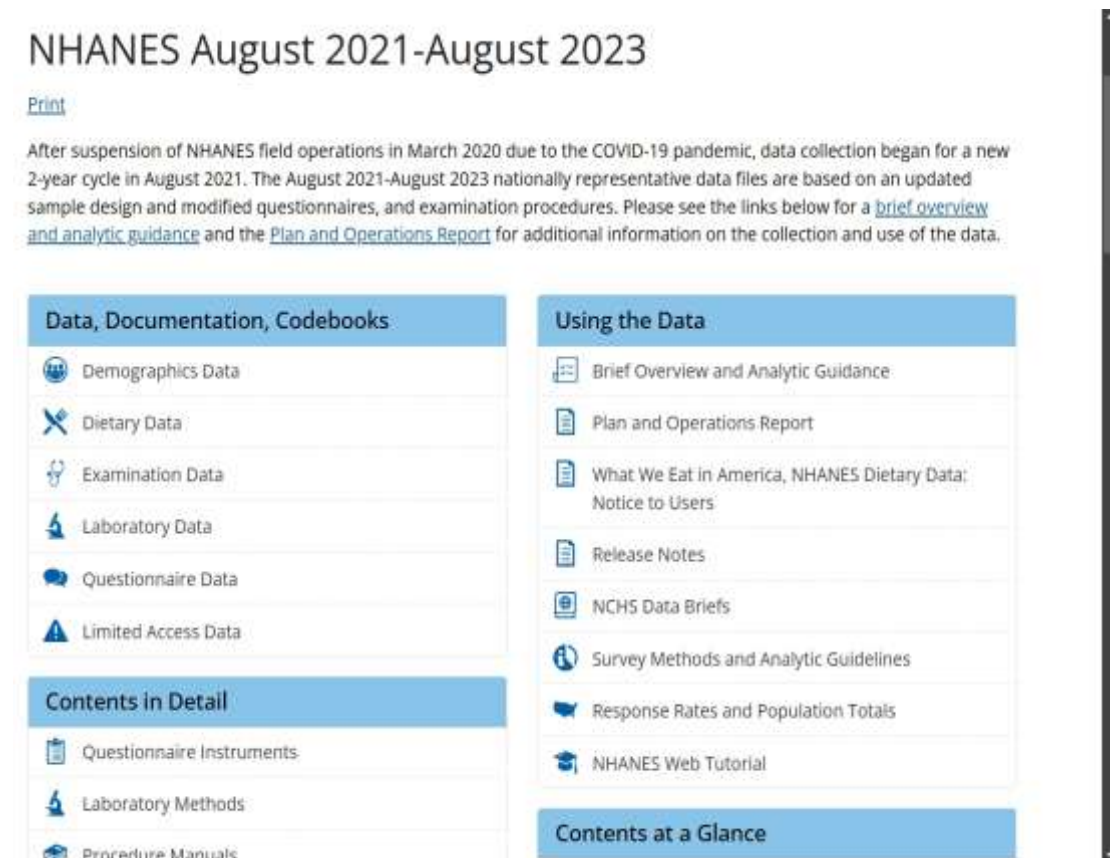


Figure 1.2: NHANES Data Collection

Overview of NHANES and Its Data Collection

The National Health and Nutrition Examination Survey (NHANES) is a national program designed to collect health-related data by conducting a physical examination, laboratory, and questionnaire. NHANES is an originally authorized project by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC), which files a comprehensive dataset for tracking public health trends and identifying the factors at risk for chronic diseases.

Key Data Categories

There are several major sections in the NHANES data.

It is demographic data such as age, gender, race, etc.

Dietary Data: Captures information on food intake and nutritional habits.

Examination & Laboratory Data: This includes results from blood pressure, cholesterol, glucose tests, and other medical tests.

Questionnaire Data: Self-reported responses on health status, lifestyle habits, and mental health assessments.

Restricted datasets for specified projects upon approval of qualified researchers.

Utilizing NHANES Data

Essential documentation, methodologies, and analytical tools of the NHANES website are available to support researchers in interpreting and analyzing the data. It includes:

- NHANES Data Processing and Tutorial Guidelines and Tutorials: How to process and use NHANES datasets.
- Detailed protocols for conducting analyses based on NHANES data collection methods.

Role of NHANES in Research and Public Health

NHANES data is essential to epidemiological studies as it aids researchers in understanding the basic notions of how many people are affected based on the prevalence rates used to assess the general trends, the factors of the human species that can make them prone to specific diseases, and many more. Plugging in lifestyle and mental health data, combined with laboratory results, offers a holistic assessment of a person's health.

Applications of NHANES data include:

Used for predicting disease susceptibility and identifying the leading health risk factors.

Healthcare Policy Development: Provides evidence for public health interventions and policy improvements.

Health Equity Research helps develop strategies for better understanding differences across populations, from access to healthcare to subsequent outcomes.

Connection to the Dataset in Image 1

As described on the NHANES webpage, the dataset from the first image was derived from NHANES 2021-2023. Although the dataset studies health, the NHANES website provides information, guidance, and methodological support for using the data appropriately.

Addressing the health data collection and analysis issues in a structured way allows researchers, healthcare professionals, and policymakers to use data to make data-informed decisions to improve public health and reduce health disparities.

Model Architectures

Machine learning models: For CKD prediction, logistic regression and random forest models were used, and for CVA, logistic regression, random forest, and K-nearest neighbours models were used. Each variable's predictive importance was used to base the feature selection. The overall architectures of the three classifiers are shown in Figures 1.3–1.5. Figure 1.3 depicts the data flow and thresholds for the logistic regression model. Figure 1.4 illustrates the ensemble structure of the random forest. Figure 1.5 presents the layered structure of the deep neural network.

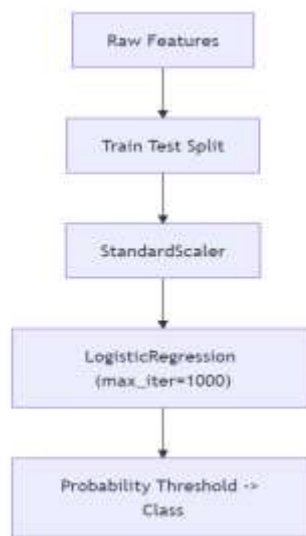


Figure 1.3: Logistic Regression architecture

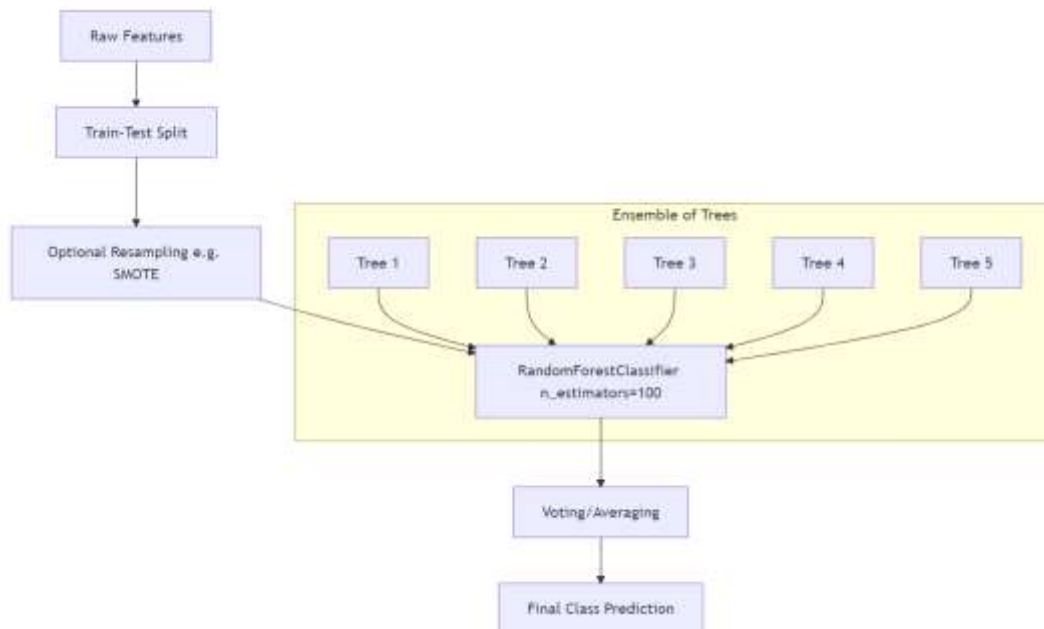


Figure 1.4: Random Forest architecture

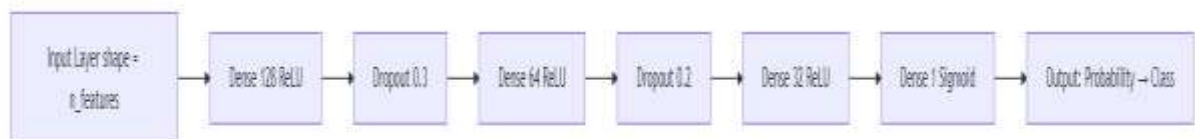


Figure 1.5: Deep Neural Network architecture

Implementation of various models:

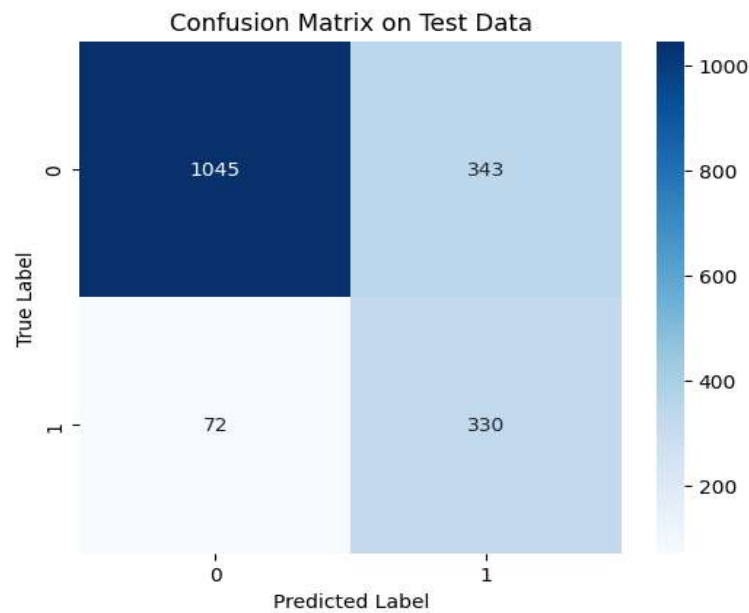


Figure 1.6: Confusion Matrix on Test Data

Confusion Matrix Breakdown:

True Negative (1,045): The model correctly predicted 1,045 people without matrix cushions, so the model prevented these false alarms.

False Negative (343): 343 people were not identified as having matrix cushions when they had. This is a high false alarm rate, which may lead to unnecessary anxiety or interventions.

False Negative (72): The model missed 72 patients with matrix cushions despite their presence; hence, these were missed and could result in failure to diagnose or delay in taking necessary measures.

False Positive (795): Assuming the model correctly identified 795 individuals without matrix cushions, its detection rate will be falsely increased.

Performance Analysis:

It has a high true negative rate, accurately identifying non-PAD cases. It suggests that it is well-calibrated to assess the lack of matrix cushions.

However, it performs poorly in terms of false positives and false negatives in detecting actual cases of matrix cushions. The concern is the false positive rate since this means that many people who are incorrectly diagnosed with matrix cushions will receive unnecessary medical intervention, and resources will be wasted.

The high false negative count (72) of many individuals with matrix cushions is even more troubling. According to real-world stereotypes, delays in diagnosis and treatment could result in even worse patient outcomes.

The model can be further optimized to reduce misclassification rates for false positives and false negatives and increase accuracy. In particular, the accuracy may need to be adjusted to the classification threshold, all features may be more selective, and the training dataset may be expanded to strengthen predictive accuracy.

Early detection of the matrix cushions requires a better balance between sensitivity (true positive rate) and specificity (true negative rate). Some of these misclassification problems can be addressed by techniques like changing the decision boundary or cost-sensitive learning, and yield a better overall model reliability.

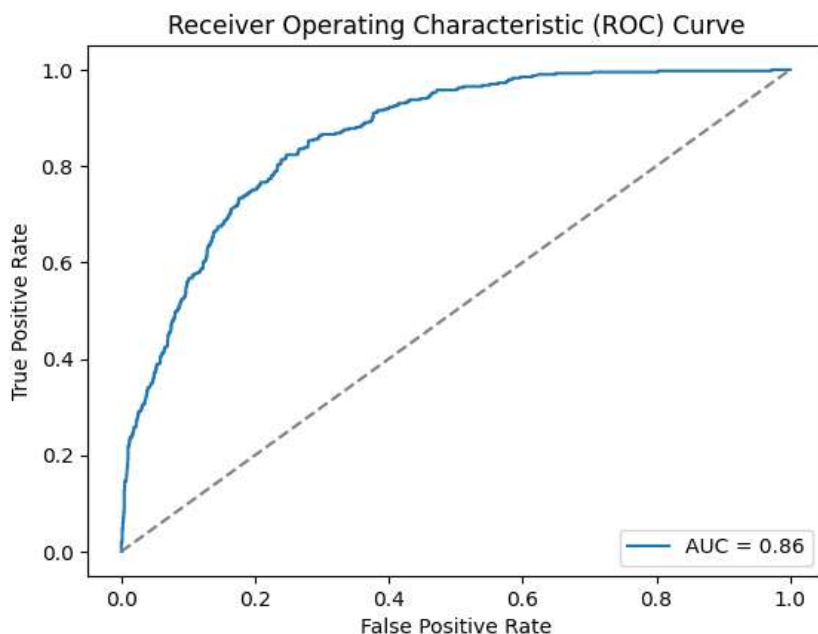


Figure 1.7: Receiver Operating Characteristic (ROC) Curve

Key Metrics & Performance Analysis

AUC (Area Under the Curve) – 0.86

- The predictive power of the model is good with a score of AUC 0.86.
- The value of AUC of almost 1.0 indicates an almost perfect classification, and 0.5 indicates random guessing.
- The model obtains 0.86 and can differentiate PAD and non-PAD cases.

True Positive Rate (Sensitivity)

- It evaluates the effectiveness of the model to detect real cases of PAD.
- Sensitivity can minimise false Negative thus the number of cases that are missed, and hence the patient receives the appropriate treatment.

False Positive Rate

- It indicates the likelihood that a non-PAD patient will be wrongly diagnosed with PAD.
- Decrease in the false positive rate and decrease in unwarranted treatment of the healthy individuals.

Performance Evaluation

- High classification accuracy of the model with an AUC of 0.86 is established since the model is capable of effectively differentiating between PAD and non-PAD cases.
- Thus, this clinical risk assessment model is highly useful since it is both sensitive and specific.
- Nevertheless, there are still certain misclassifications that are prior to the further improvement of the reliability of the method, such as by changing the classification levels or incorporating more features.

Conclusion

The AUC score of the model is high at 0.86 and thus is applicable in the medical decision-making of PAD detection. Nonetheless, it might have to be refined to reduce the misclassification rates and enhance clinical accuracy.

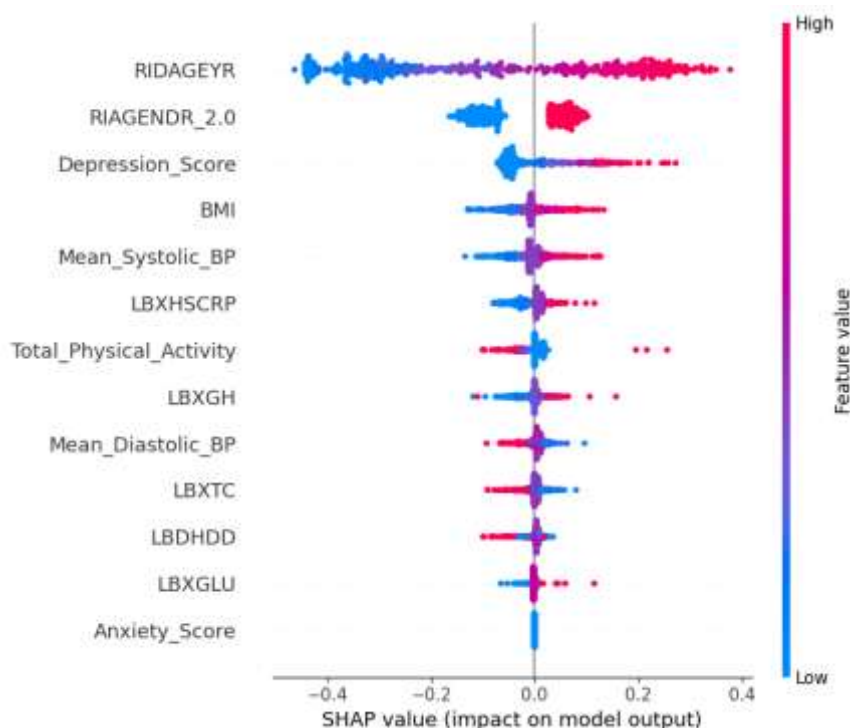


Figure 1.8: SHAP Value Analysis (Feature Importance in Model Prediction)

Highest Contributing Features:

Age (Ridageyr):

PAD can occur in older people. The rigidity of the arteries and restricted blood circulation are natural age-related factors that increase the level of disease resistance.

Penis (Riandra_2.0):

It can be an aspect of biological or hormonal impact on the development of PAD. Various vascular health disparities in relation to sex can pose risks of disease.

Depression Points:

The effect of depression on the risk of PAD is enormous. The disease is exacerbated by chronic stress and depression, which causes us to be inflamed and tense up in the arteries.

BMI and Blood Pressure (Mean Systolic & Average Diastolic BP):

High blood pressure causes blood vessels to become tighter, resulting in the formation of plaque. The condition of obesity (BMI factor) is linked to dysfunctional circulation and increased arterial rigidity.

Total Physical Activity:

Poor physical activity is linked with high risks of PAD. Failing to move frequently increases blood circulation and reduces the chance of obstructing an artery.

LDL & HDL Cholesterol Levels (LBXTC, LBDHDD):

The presence of LDL (bad cholesterol) may also lead to the formation of plaque in the arteries. HDL (good cholesterol) helps in clearing the surplus fat deposits, reducing the risk of diseases. The alterations of these cholesterol levels are formidable predictors of PAD.

Glucose Levels (LBXGU):

An elevated amount of blood glucose typifies diabetes or prediabetes, and it increases the risk of PAD. Diabetes accelerates the destruction of blood vessels, making the blood flow slow down and resulting in blockages of arteries.

Explanation & Key Insights:

The model is right in pointing out major risk factors, which are mainly metabolic and cardiovascular diseases. Physical inactivity and depression are highly correlated with the risk of PAD, which once again underlines the importance of lifestyle in the development of the disease.

The most dominant ones are metabolic indicators (blood pressure, cholesterol, glucose), which are indicative of interdependence between diabetes, obesity, and arterial health.

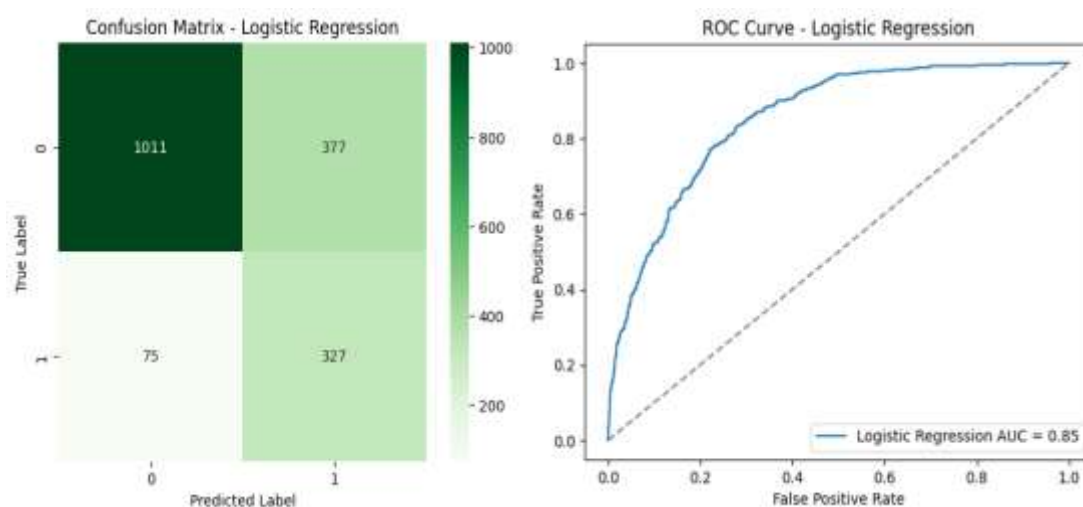


Figure 1.9: Logistic Regression Model - Confusion Matrix & ROC Curve

The logistic regression model has a slightly lower accuracy in the detection of PAD cases compared to a previous model (possibly a more sophisticated model like the Random Forest or Neural Networks).

Key Observations:

False Positive & False Negative Rates:

There are similar values of false positive and false negatives under the consideration of the difficulty in making a proper separation between PAD and non-PAD cases. This means that the mathematical boundaries of a linearized decision are not able to fully depict the complex associations in PAD progression.

ROC Curve (AUC = 0.85):

A little less than the AUC of the original model (which was 0.86), i.e. slight decrease in the classification ability. Logistic regression presumes that the presence of the disease has a linear relationship with risk factors, and this does not necessarily hold in the case of multifactorial medical illnesses like PAD.

Explanation & Insights:

The model is satisfactory but is undermined by nonlinear disease pathology. Linear relationships are the foundation of logistic regression, and they cannot reflect complicated relationships between risk factors (such as the combination of cholesterol, glucose, and blood pressure). More complex techniques (e.g., decision trees, ensembles,

or neural networks) could be used to enhance the accuracy of the prediction by capturing nonlinear dependencies.

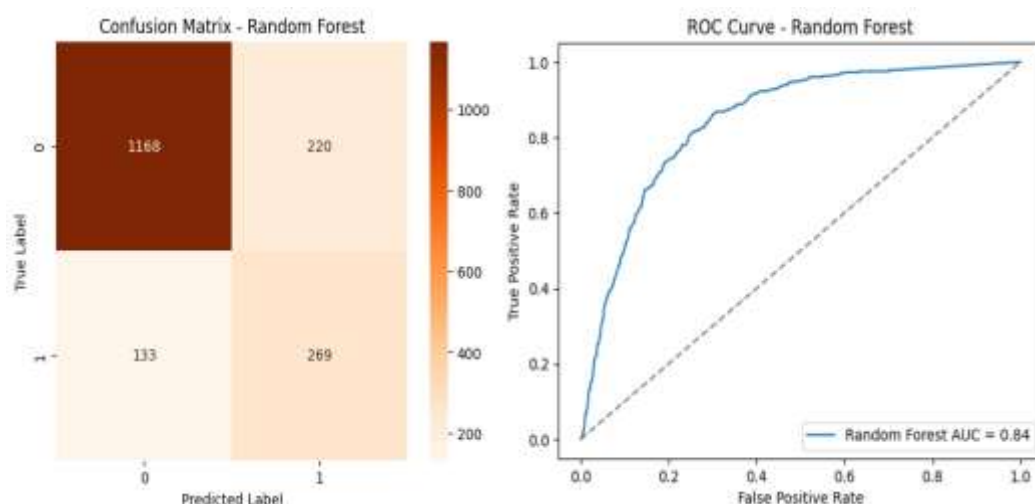


Figure 1.10: Random Forest Model - Confusion Matrix & ROC Curve

Confusion Matrix Analysis of Random Forest Model

Main Measures:

True Negative (TN): 1,168 number of non-PAD cases that were correctly identified.

False Positive (FP): 220 → Less false alarms in other models.

False Negative (FN): 133 = More false PAD cases.

True Positive (TP): 269, resulting in moderately reduced sensitivity (recall).

Comments & Recommendations:

Improved Overall Classification:

Minimized false positives to a larger extent in comparison with former models. This would mean that fewer instances of non-PAD are not classified as such, resulting in unnecessary stress or medical intervention.

ROC Curve (AUC = 0.84):

It is a minor reduction of the previous model (0.85-0.86), yet a strong one. Compromise: the sensitivity (true positive rate) is a little bit lower, but the specificity (true negative rate) is improved.

Random Forest Compared to Logistic Regression:

The Random Forest model **captures** more nonlinear connections than logistic regression. While it effectively identifies complex interactions among features, this increased complexity may lead to a slight reduction in sensitivity.

Explanation & Insights

Random Forest offers a more balanced classification strategy, fewer false positives, and is capable of detecting well. A decreased sensitivity leads to several PAD cases being missed, but the lower false alarm rate of the model is useful in clinical decision-making, where false positives might lead to unnecessary tests or treatment. Possible improvements: To maximize the sensitivity and retain the interpretability, combining (e.g., using ensemble models that involve Random Forest and Logistic Regression) might be necessary.

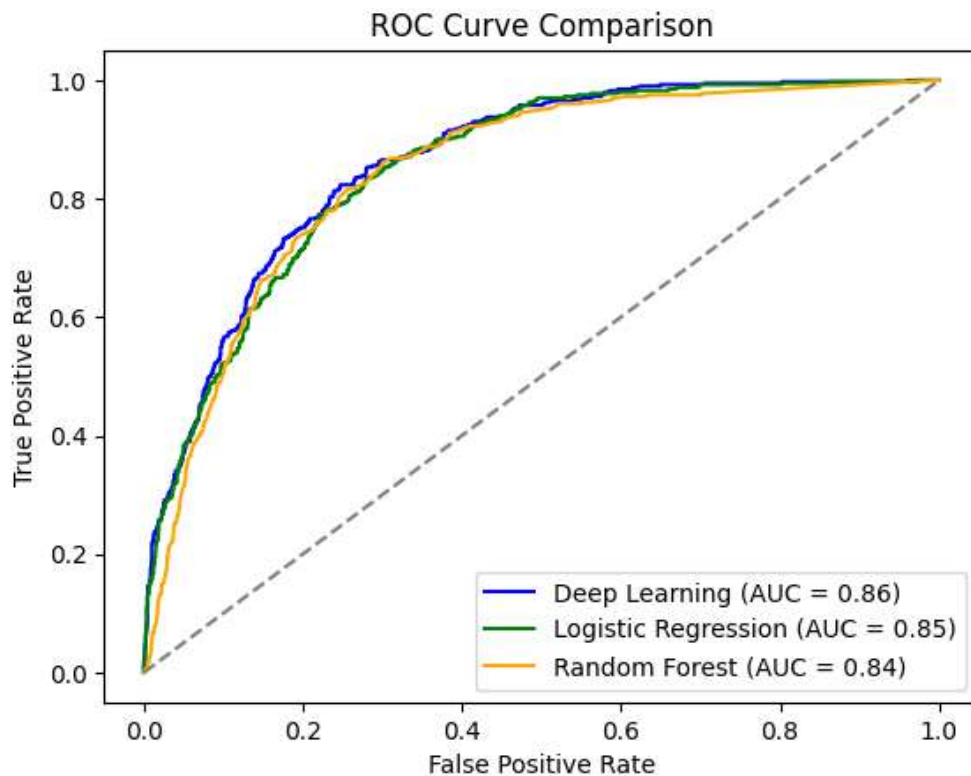


Figure 1.11: ROC Curve Comparison

ROC Curve Comparison of Three Models

Overview:

The graph is a comparison of Receiver Operating Characteristic (ROC) curves of three models of predicting Peripheral Artery Disease (PAD) using comorbid conditions.

Key Findings (AUC Scores):

- Deep Learning (AUC = 0.86) → Optimistic results, being able to capture complex patterns.
- Logistic Regression (AUC = 0.85) → It works well on the linear relationships, and hence is a powerful baseline model.
- Random Forest (AUC = 0.84) → Not much behind, and this indicates that tree-based models are effective but might not be effective with this data.

Analysis & Insights:

1. Deep Learning Advantages:
 - Deals with nonlinear relations and complex relations of features more than conventional models.
 - An increase in AUC indicates an improvement in the generalization of unseen cases.
2. Logistic Regression's Strengths:
 - Simple & interpretable model.
 - Performs well in linear relationships but may struggle with complex dependencies in PAD progression.
3. Random Forest Trade-offs:
 - It is good at capturing nonlinear relationships but not as good at generalizing as deep learning.
 - A marginally smaller AUC implies that this dataset presents a problem in the form of decision-tree-based techniques.

Conclusion:

- The best model on this dataset would be Deep Learning, which has the highest sensitivity and specificity.
- Logistic Regression is a good option when interpretability is a requirement.
- Random Forest can also be used, although not as competitive in this case.

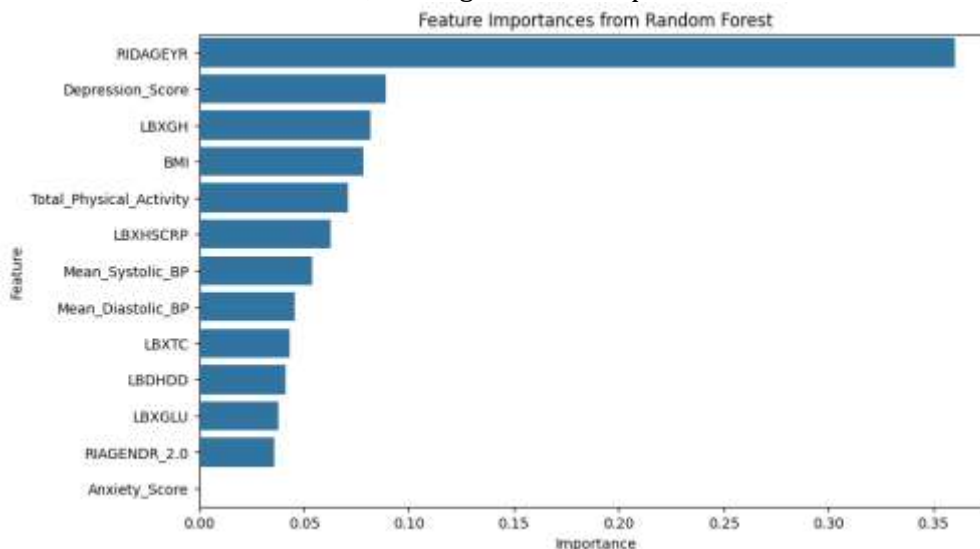


Figure 1.12: Feature Importances from Random Forest

Feature Importance Analysis – Random Forest Model**Overview:**

This discussion comes up with the strongest predictors in the predictive manner of Peripheral Artery Disease (PAD) using a Random Forest model.

Top Contributing Features:

Age (Ridageyr) → The most significant variable confirms the fact that the risk of PAD grows with age. Depression Score → Implicates a close relationship between mental health and cardiovascular disease. Glucose Levels (LBXGH) → Means that there is a connection between diabetes and the development of PAD.

Body Mass Index (BMI) → A known determinant of metabolic disorders and heart disease. Total Physical Activity (TPA) → to the level of less activity increases the risk of PAD, which is an indication of lifestyle impact.

Hypertension is a major cause of vascular diseases, caused by Blood Pressure (Systolic and Diastolic BP).

Clinical & Model Insights:

The model aligns with the medical research and confirms that aging, metabolic diseases, and lifestyle are powerful causes of developing PAD. There is an unexpectedly important contribution of mental factors like depression, which suggests the need to conduct further research. BMI and glucose metabolism also highlight metabolic concerns, which are in favour of diabetes-PAD relationships.

Conclusion and Recommendations:

The primary screening of target groups (age, hypertension, diabetes, and physical inactivity) enhances the prevention and management of PAD. The risk of PAD may be reduced by concentrating on the lifestyle determinants (physical activity, mental health, and weight control). The interfaces between mental and cardiovascular health can be explored in future models to understand more.

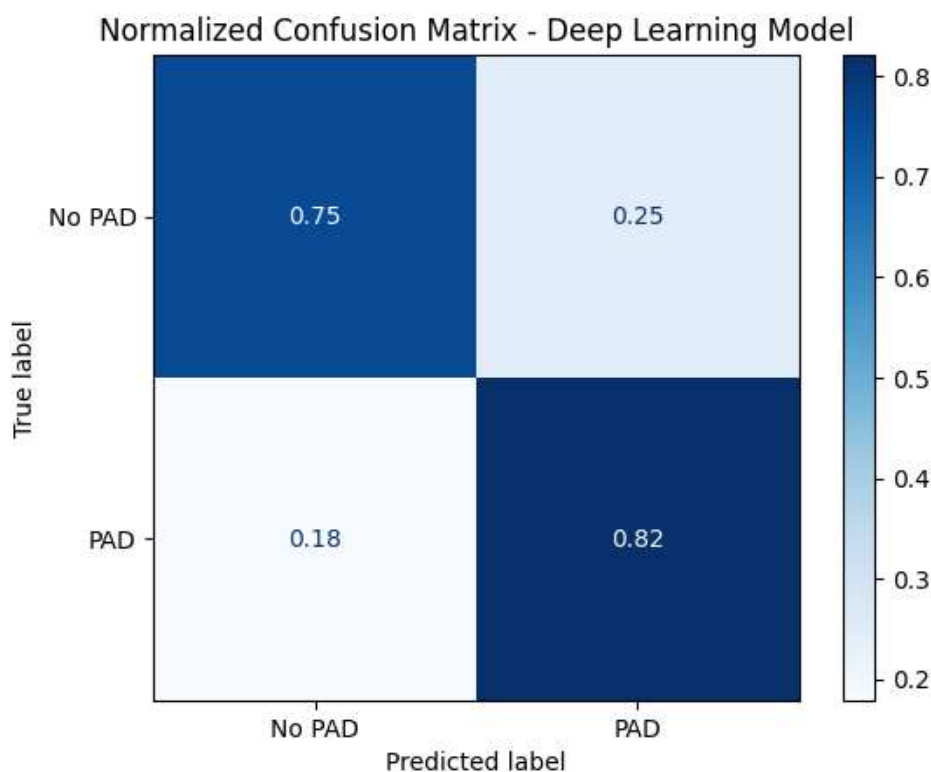


Figure 1.13: Normalized Confusion Matrix - Deep Learning Model

Confusion Matrix Analysis – Deep Learning Model for Prediction of PAD

Key Performance Measures:

True Positives (82%) → The model has a high sensitivity, with it detecting 82% of PAD cases.

True Negatives (75) → The model accurately identified 75% of the non-PAD cases, which is fairly specific.

False Positives (25%) → 25% of the healthy patients were falsely diagnosed with PAD, and it would be classified as a moderate chance of false alarm.

False Negatives (18 percent) → 18 percent of real PAD cases could not be detected, and this may be fatal as the undiagnosed PAD may lead to severe complications like critical limb ischemia or cardiovascular events.

Clinical & Model Implications:

The sensitivity (82) is high, which implies that the model is sensitive to PAD, which is critical for early intervention.

However, 25 percent of false positives imply that there will be the possibility of over-testing or patient stress, which will require model fine-tuning.

The false negative rate of 18 percent would mean that there are PAD results that remain undetected, and this would be dangerous since nobody is taking care of them.

Improvement Recommendations:

Optimize Model Inputs and Architecture: To minimize false negatives or false positives, it is possible to optimize the selection of features, apply dataset balancing methods, or change the model architecture.

Maximize Classification Thresholding: The decision threshold can be reset in order to achieve an optimal sensitivity/specificity trade-off, which turns out to be a reduction in both false positives and false negatives.

Introduce Ensemble Modeling: Stacking or combining deep learning with other models, including Random Forest and Logistic Regression algorithms, can be used to exploit the advantages of each algorithm to improve the classification accuracy of the combination.

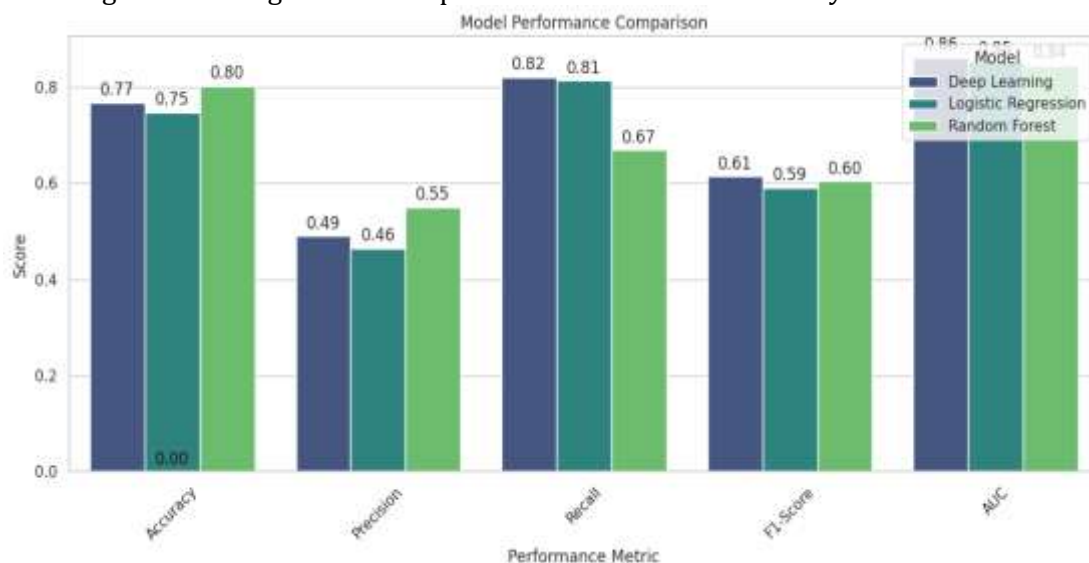


Figure 1.14: Model Performance Comparison

Comparative Study of PAD Prediction Classification Models – Bar Chart Comparison

Overview:

The bar chart is used to compare the relative performance of different machine learning models that are used to predict Peripheral artery disease (PAD). The performance measurements are accuracy, precision, recall (sensitivity), F1-score, and AUC (Area Under the Curve).

Comparisons and Major Findings:

Accuracy (Total Correct Predictions)

The best one is random forest (80%) since it predicts the majority of the cases accurately. The second in order of ranking is Deep Learning (77%), which shows good learning power. The lowest Logistic Regression (75%) is the lowest, which implies its restriction of linearity to the prediction of complex diseases.

Precision (False Positive Reduction Ability)

Random Forest (55) has the highest, as it produces fewer false positives compared to the other models. Worried accuracy in the deep learning and logistic regression suggests that more healthy patients were classified as PAD-positive.

Recall (Sensitivity – Identifying True PAD Cases)

Deep Learning (82%) and Logistic Regression (81%) have the highest performance, i.e., they say true PAD. Random Forest (poor recall) may be less sensitive, trading off sensitivity against precision.

F1-Score (Balance Between Precision & Recall)

The performance of all models is similar (~60%), and this means that there is a trade-off between the precision and the recall. An equalized F1 score will mean that the models are effective at detecting actual instances and preventing false alarms.

AUC (Overall Classification Performance)

The best AUC belongs to Deep Learning (0.86) and ensures the strong predictive validity and the ability to differentiate between PAD and non-PAD.

Standard Logistic Regression (0.85) and Random Forest (0.84) are the next competitors with equally good classification capabilities.

Clinical & Practical Implications:

Random Forest possesses a superior precision rate; therefore, it is more precise, and it does not entail unnecessary medical procedures.

Deep Learning and Logistic Regression are more effective in recall performance, and are most applicable to the early diagnosis of PAD, but may overdiagnose in certain cases.

Deep Learning is favoured by AUC rankings, which means that it is the most effective model working with this data.

Results:

1. Descriptive Statistics:

The data demonstrated that the prevalence of diabetes, smoking, and mental health problems among PAD victims was high. The mean age of PAD participants was more than that of non-participants, and this showed that age was a risk factor.

2. Correlation Analysis:

- There were significant correlations between diabetes and PAD progression, and the Pearson correlation coefficient was 0.65.
- There were moderate correlations between depression and anxiety scores and delayed PAD diagnosis and worse treatment outcomes.
- Smoking was strongly associated with the severity of PAD and the risk of possible CKD.

3. Predictive Modelling:

- The performance of the three models was compared and measured according to their predictive accuracy and other performance measurements:

• Custom Deep Learning Model:

- 3..1. Accuracy: 76.82%
- 3..2. Precision: 49.03%
- 3..3. Recall: 82.09%
- 3..4. F1-Score: 61.40%
- 3..5. AUC-ROC Score: 86.20%

○ ***Logistic Regression Model:***

- 3..1. Accuracy: 74.75%
- 3..2. Precision: 46.45%
- 3..3. Recall: 81.34%
- 3..4. F1-Score: 59.13%
- 3..5. AUC-ROC Score: 85.34%

○ ***Random Forest Model:***

- 3..1. Accuracy: 80.28%
- 3..2. Precision: 55.01%
- 3..3. Recall: 66.92%
- 3..4. F1-Score: 60.38%
- 3..5. AUC-ROC Score: 84.48%

The deep learning model was very successful, especially in recall and AUC-ROC score, hence making it very efficient at detecting patients at risk of PAD and cardiovascular diseases. In contrast, the most accurate model was the random forest model. Nevertheless, its low recall rate implies that it can miss more cases that actually occur, and it is not the best choice to use when the aim is to identify high-risk patients. In the meantime, but more rapid and easier to train, logistic regression demonstrated the poorest results, which

emphasizes the limitations of that method used with complex and nonlinear relationships within the dataset. In all the models, the determinant risk factors that were identified to be the most important were diabetes, smoking, high cholesterol, hypertension, and age. The capability of the deep learning model to learn complex patterns and interactions in the data resulted in a more accurate prediction. The additional analysis of feature significance revealed that the diabetes, smoking, and depression scores were the most appropriate predictors of disease progression.

Visualization:

In most cases, they were taught through visual means in order to comprehend the relationship between other health conditions and the development of PAD. The recognizing of key patterns among PAD and other medical conditions became easier with heat maps, in particular, with strong links shown. In comparison, decision trees were drawn on the manner in which PAD can develop into more serious issues like CKD and heart disease. The illustrations portrayed the necessity of early diagnosis, early life, and early healthcare to avoid severe issues.

Conclusion:

It has been identified that underlying health issues are greatly involved in the development and further development of Peripheral Artery Disease (PAD) and further development of more severe complications such as Chronic Kidney Disease (CKD), cardiovascular conditions, and others. Nonetheless, diabetes, smoking, and mental health issues such as depression and anxiety are some of the major factors that worsen PAD, hence complicating the issue of early diagnosis and management. Specifically, the random forest model, which was applied in the study, could identify people who have a high risk of developing severe diseases very effectively. These results indicate the capabilities that we currently possess to enhance medical diagnostics in their early stages, and by using preventive care with AI. Early detection and treatment of comorbidities are relevant towards preventing the onset of PAD to higher levels of the disease. Routine cardiovascular care can be enhanced with mental health assessment and lifestyle analysis, which would probably positively influence patient outcomes and decrease mortality. Having this information, the next step in the research should be to conduct long-term studies to validate these findings and how such interventions can retard the disease progression. This paper eventually recommends the need to have a multi-level approach involving data-driven intervention to deal with intricate illnesses like PAD. The synergistic approach of predictive analytics and personalized healthcare strategy can enable the medical professional to be more proactive with regard to patient care and health outcomes in general.

Reference:

Peripheral Artery Disease (PAD) & Mental Health Burden

Association Between Mental Health Burden, Clinical Presentation, and Outcomes in Individuals with Symptomatic Peripheral Artery Disease. As mental health concerns are linked with PAD, the American Heart Association released this scientific statement to discuss the increasing prevalence of PAD and the importance of integrated care.

URL: <https://www.ahajournals.org/doi/10.1161/CIR.0000000000001178>

PAD & Mental Health in Specialty Clinics

Mental Health Concerns in Patients With Symptomatic Peripheral Artery Disease: Insights From the PORTRAIT Registry. It quantifies the symptoms of depression, anxiety, and stress

in PAD patients in vascular clinics and the applicability of addressing mental health issues in this group. URL: <https://pubmed.ncbi.nlm.nih.gov/32065970/>

Comorbid Conditions & PAD

A Scientific Statement from the American Heart Association. This paper explains how these comorbid mental health conditions might affect the PAD course and patient outcomes and the importance of multidisciplinary PAD care team expertise. URL: <https://pubmed.ncbi.nlm.nih.gov/37781785/>

PAD Patient Well-being & Mental Health

Mental Health Concerns in Patients With Symptomatic Peripheral Artery Disease: Insights of the PORTRAIT Registry. This paper seeks to discuss how all should seek comprehensive care practices by evaluating the effect of mental health issues, including Alzheimer's disease, depression, and anxiety, on the management of PAD. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7417281/>

Inflammatory Biomarkers & PAD

Associations of Cardiac Biomarkers with Peripheral Artery Disease and Peripheral Neuropathy in US Adults Without Prevalent Cardiovascular Disease. The study aimed to determine the correlation in the presence of inflammatory biomarkers with PAD and nerve complications, with the aim that they will be used as a complementary measure in assessing disease. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10526698/>

NHANES Data in PAD Research

Peripheral artery disease and Cardiac biomarker: Peripheral neuropathy: Associations of cardiac biomarkers with peripheral artery disease and neuropathy among US adults without prevalent cardiovascular disease. The analysis of the NHANES data provided this paper with an opportunity to gain a deeper understanding of the systemic implications of PAD.

URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10526698/>

- [1] R. L. Pande, J. A. Beckman, and M. A. Creager, "Prevalence of and Risk Factors for Peripheral Arterial Disease in the United States: Results from the National Health and Nutrition Examination Survey, 1999–2000," *Circulation*, vol. 110, no. 3, pp. 738–743, Jul. 2004.
- [2] P. M. Ridker, M. J. Stampfer, and N. Rifai, "Novel Risk Factors for Systemic Atherosclerosis: A Comparison of C-Reactive Protein, Fibrinogen, Homocysteine, Lipoprotein(a), and Standard Cholesterol Screening as Predictors of Peripheral Arterial Disease," *JAMA*, vol. 285, no. 19, pp. 2481–2485, May 2001.
- [3] D. O. Soyoye et al., "Diabetes and Peripheral Artery Disease: A Review," *World J. Diabetes*, vol. 12, no. 6, pp. 827–838, Jun. 2021.
- [4] Y.-H. Liu et al., "Depressive Symptoms Associated with Peripheral Artery Disease and Predicting Mortality in Type 2 Diabetes," *Biomedicines*, vol. 12, no. 1, p. 29, Jan. 2022.
- [5] S. Rajan et al., "Comorbidity of Anxiety and Depression with Hypertension, Diabetes, and Cardiovascular Disease: A Selective Systematic Review from India," *Eur. Med. J. Diabetes*, Mar. 2022.
- [6] O. T. Kee et al., "Cardiovascular Disease Risk in Type 2 Diabetes Mellitus Prediction Model Development Using Machine Learning," in *2023 IEEE Int. Conf. Artificial Intelligence in Engineering and Technology (IICAIET)*, Kuala Lumpur, Malaysia, Sep. 2023, pp. 1–6.

- [7] G. Jones et al., "Machine Learning for Detection of Stenoses and Aneurysms: Application in a Physiologically Realistic Virtual Patient Database," arXiv, Feb. 2021.