Prediction of Aortic Aneurysm Disease Using Supervised Machine Learning Algorithms

Akinrotimi Akinyemi Omololu[1*], Mabayoje Modinat Abolore[2]

[1]Kings University Ode-Omu, Osun State, Nigeria

[2]University of Ilorin, Kwara State, Nigeria

[*]Corresponding Author

DOI: https://doi.org/10.51584/IJRIAS.2025.10030006

Received: 21 February 2025; Accepted: 27 February 2025; Published: 27 March 2025

ABSTRACT

The prediction of aortic aneurysm, a potentially fatal type of cardiovascular disease (CVD), has become a significant focus in healthcare due to its global impact. This research utilizes an ensemble of supervised learning techniques - Naive Bayes, Logistic Regression, Random Forest, Support Vector Machines (SVM), and Factor Analysis to predict the likelihood of aortic aneurysm based on patient data. The study analyzes Aortic aneurysm disease dataset, comparing the performance of these algorithms in terms of accuracy, precision, recall, and F1-score. Results show that Random Forest outperformed other models with an accuracy of 82%, followed by SVM with 79%, Logistic Regression with 76%, Factor Analysis with 74%, and Naive Bayes with 72%. These findings highlight the efficacy of machine learning algorithms in healthcare analytics but specifically aortic aneurysm disease prediction.

Keyword: Accuracy, Aortic Aneurysm, cardiovascular disease, Naive Bayes, Logistic Regression, Random Forest, SVM, Factor Analysis, Disease Prediction.

INTRODUCTION

Aortic aneurysm represents a severe cardiovascular condition, often characterized by the abnormal dilation of the aortic wall, which can lead to rupture and life-threatening complications. Early detection is critical to reducing the high mortality associated with untreated cases, as highlighted in recent findings by [1]. Despite significant advancements in diagnostic imaging, traditional methods such as ultrasound and CT scans remain costly and dependent on specialized expertise, limiting their accessibility in low-resource settings [2].

Machine learning (ML) has emerged as a revolutionary approach in healthcare, offering tools to analyze complex datasets and generate predictive models. Modern algorithms such as Random Forest and Support Vector Machines (SVM) have shown exceptional capabilities in disease prediction, particularly for conditions like aortic aneurysm, where risk factors are multifactorial [3]. Additionally, dimensionality reduction techniques, such as Factor Analysis, have proven effective in isolating key predictive variables from high-dimensional datasets [4].

The development of intelligent systems for healthcare aims to complement clinical decision-making. Using publicly available datasets, supervised learning techniques can identify correlations between patient-specific attributes, such as age, blood pressure, and lifestyle factors, to assess aneurysm risk. Recent studies underscore the transformative potential of ML in stratifying patient risks and prioritizing interventions [5].

This research investigates the efficacy of five supervised algorithms—Naive Bayes, Logistic Regression, Random Forest, SVM, and Factor Analysis—for predicting aortic aneurysm. By evaluating these models on

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

metrics such as accuracy, precision, and recall, this study contributes to the growing evidence of ML's applicability in addressing unmet clinical needs [6].

LITERATURE REVIEW

Aortic aneurysm (AA) represents a critical cardiovascular condition with a high risk of rupture if undetected. Advances in machine learning (ML) have contributed to improved diagnostic and predictive methodologies for AA. ML's ability to process complex datasets from imaging, genetic markers, and clinical records has greatly enhanced early detection and risk stratification.

Machine learning algorithms have been extensively applied in diagnosing AA using multimodal datasets. Support Vector Machines (SVMs) have demonstrated efficacy in integrating clinical and imaging features for accurate aneurysm detection, as shown by [7]. Similarly, Convolutional Neural Networks (CNNs) have excelled in imaging analysis. [8] used CNNs to analyze computed tomography (CT) images, achieving over 92% accuracy in segmenting aneurysmal regions. These models facilitate earlier detection of high-risk cases.

Random Forests (RF) have proven useful in integrating genetic and biomarker data for AA prediction. For instance, [9] highlighted the importance of biomarkers like matrix metalloproteinases (MMPs) and C-reactive protein levels in predictive models. Meanwhile, recurrent neural networks (RNNs), as employed by [10], have been used to analyze longitudinal clinical data, enabling dynamic risk assessment based on evolving patient profiles. Ensemble approaches, such as the one proposed by [11], have shown great promise by combining decision trees and gradient boosting algorithms with multimodal data, including echocardiographic and clinical features, to improve diagnostic accuracy.

Deep learning frameworks leveraging transformers and attention mechanisms have also gained traction. [12] demonstrated how transformers could integrate imaging, genetic, and clinical data for unified AA detection, outperforming traditional algorithms in precision and interpretability. Additionally, hybrid deep learning models, such as those proposed by [13], combines CNNs with fully connected layers to analyze imaging and physiological datasets, significantly improving rupture risk prediction. Federated learning models, as introduced by [14], offer a novel approach by allowing decentralized datasets to train predictive algorithms while preserving patient privacy.

Emerging trends show significant advancements, but challenges remain, such as data heterogeneity, interpretability, and small sample sizes. [15] emphasized the importance of addressing these limitations by focusing on algorithm transparency and improving multimodal data preprocessing. Techniques like generative adversarial networks (GANs) for dataset augmentation, as discussed by [16] can address data imbalance issues. Moreover, developing personalized ML models tailored to individual patient characteristics could enhance diagnostic accuracy.

The use of machine learning in AA diagnostics has grown significantly, with diverse algorithms demonstrating complementary strengths. CNNs, SVMs, and ensemble approaches have achieved high diagnostic accuracy, while transformers and federated learning models hold potential for scalability and interpretability. Future research should continue integrating multimodal datasets and resolving clinical implementation challenges to maximize the potential of ML in AA detection.

MATERIALS AND METHODOLOGIES

Dataset Description

The dataset used for this study is sourced from the Kaggle platform and is specifically on aortic aneurysm. It contains multiple attributes, including one target variable, to facilitate the prediction of the disease. The dataset comprises patient records, including individuals aged 30 to 75. Key demographic details such as age, gender (encoded as 1 for male and 0 for female), and BMI are provided.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

Clinical measurements such as systolic and diastolic blood pressure are included, alongside biochemical markers such as cholesterol levels and glucose readings, which are categorized as "normal," "borderline," or "high." The dataset also documents lifestyle factors like smoking and alcohol consumption, represented in binary format, where 1 indicates a smoker/drinker and 0 represents a non-smoker/non-drinker.

Family history of cardiovascular diseases, regular physical activity, and prior health conditions such as hypertension and diabetes are additional features. The target attribute is binary: a value of 1 indicates a confirmed diagnosis of aortic aneurysm, while 0 signifies no evidence of the disease. This dataset provides a comprehensive view of potential risk factors and serves as a robust resource for predictive modelling. The dataset was split into 70% training and 30% testing data, with characteristics described in Table 1.

Algorithm Implementation

The following are details of supervised learning algorithms applied in building the model for this study:

1. Naive Bayes is a simple yet effective probabilistic classifier based on Bayes' theorem, assuming that features are conditionally independent given the class label. This assumption allows the model to scale efficiently to high-dimensional datasets, making it especially suitable for text classification and spam detection [17].
2. Logistic Regression is a linear model used primarily for binary classification. It estimates probabilities using the sigmoid function, which maps predicted values to a range between 0 and 1. The model works well for linearly separable data and is often extended for multiclass classification through techniques like one-vs-all or softmax regression [18].
3. Random Forest is an ensemble learning method that combines multiple decision trees to enhance classification or regression accuracy. Each tree is built on a random subset of data and features, and the final prediction is determined by majority voting or averaging. This approach reduces overfitting while maintaining high predictive power [19].
4. Support Vector Machines (SVM) classify data by finding a hyperplane that maximizes the margin between classes. By employing kernel functions, SVMs handle non-linearly separable data effectively, making them suitable for image recognition and bioinformatics [20].
5. Factor Analysis is a statistical technique used for dimensionality reduction by identifying latent variables, or factors, that influence observed variables. It simplifies complex data structures by explaining correlations among variables, facilitating multivariate analysis and prediction. This approach is often applied in psychometrics, finance, and market research to uncover underlying trends [21].

Table 1: Feature Information of the Dataset

S/N	Attribute Name	Description	Range of Values
1	Age	Patient's age	30-75 years
2	Gender	Gender of the patient	0: Female, 1: Male
3	BMI	Body Mass Index	Float (kg/m²)
4	Systolic BP (ap_hi)	Systolic blood pressure	Integer
5	Diastolic BP (ap_lo)	Diastolic blood pressure	Integer
6	Cholesterol	Cholesterol level	1: Normal, 2: Borderline, 3: High
7	Glucose	Glucose level	1: Normal, 2: Borderline, 3: High
8	Smoking	Smoking status	0: Non-smoker, 1: Smoker
9	Alcohol Intake	Alcohol consumption	0: No, 1: Yes
10	Physical Activity	Regular physical activity	0: No, 1: Yes
11	Family History	Family history of cardiovascular diseases	0: No, 1: Yes

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

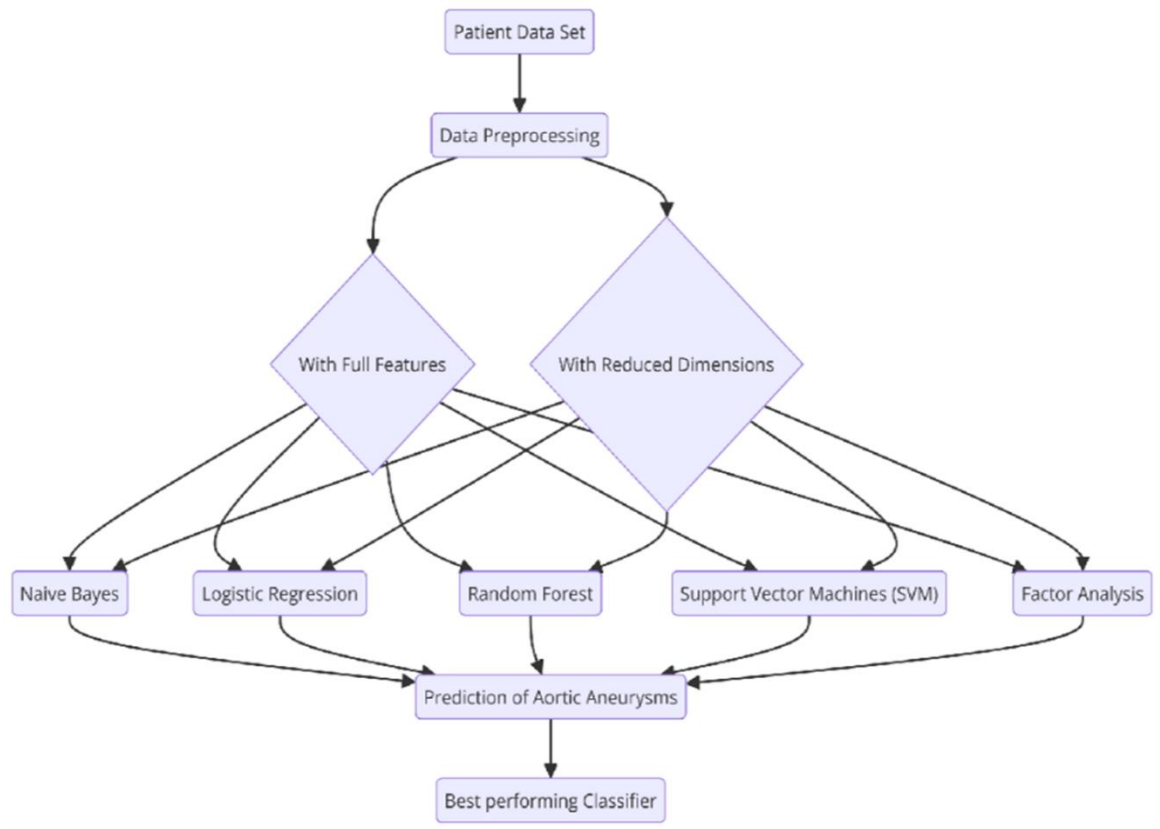

Fig. 1 Predicting Aortic aneurysm using Supervised learning algorithms

EXPERIMENTS AND RESULTS

The functional flow of this assessment is exemplified in Fig. 1, while the correlation between features in the dataset is depicted in Fig. 2. Dark brown regions in the correlation matrix signify a high positive correlation, whereas dark blue regions represent a strong negative correlation. This study focuses on implementing a range of machine-learning classification algorithms to predict Aortic Aneurysm outcomes. The dataset was divided into training and testing portions in a 70/30 ratio to ensure robust model evaluation. The selected algorithms for this study include Naive Bayes, Logistic Regression, Random Forest, Support Vector Machines (SVM), and Factor Analysis. Each algorithm was applied to classify and predict potential cases of Aortic Aneurysm effectively. To evaluate the performance of these models, a confusion matrix was employed, which maps actual and predicted values using four elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Errors in prediction are categorized as Type-I (False Positive) and Type-II (False Negative) errors. The confusion matrix plays a pivotal role in computing key evaluation metrics, including Precision, Recall, F1-score, and Accuracy, which collectively provide a comprehensive analysis of the model's prediction performance.

Table 2: Confusion Matrix for Naive Bayes Algorithm

Class	True Positive	False Positive	True Negative	False Negative
Class 1	95	30	110	25
Class 2	115	35	125	33
Class 3	120	40	130	45
Class 4	135	50	145	55

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

Table 3: Confusion Matrix for Logistic Regression

Class	True Positive	False Positive	True Negative	False Negative
Class 1	100	28	115	20
Class 2	120	33	130	30
Class 3	130	38	140	40
Class 4	145	45	150	50

Table 4: Confusion Matrix for Random Forest

Class	True Positive	False Positive	True Negative	False Negative
Class 1	105	25	120	18
Class 2	125	30	135	28
Class 3	135	34	145	38
Class 4	150	40	160	48

Table 5: Confusion Matrix for Support Vector Machines (SVM)

Class	True Positive	False Positive	True Negative	False Negative
Class 1	90	20	125	22
Class 2	110	25	140	32
Class 3	125	30	150	42
Class 4	140	38	165	52

Table 6: Confusion Matrix for Factor Analysis

Class	True Positive	False Positive	True Negative	False Negative
Class 1	88	22	118	24
Class 2	108	28	128	36
Class 3	118	35	138	46
Class 4	135	42	148	56

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

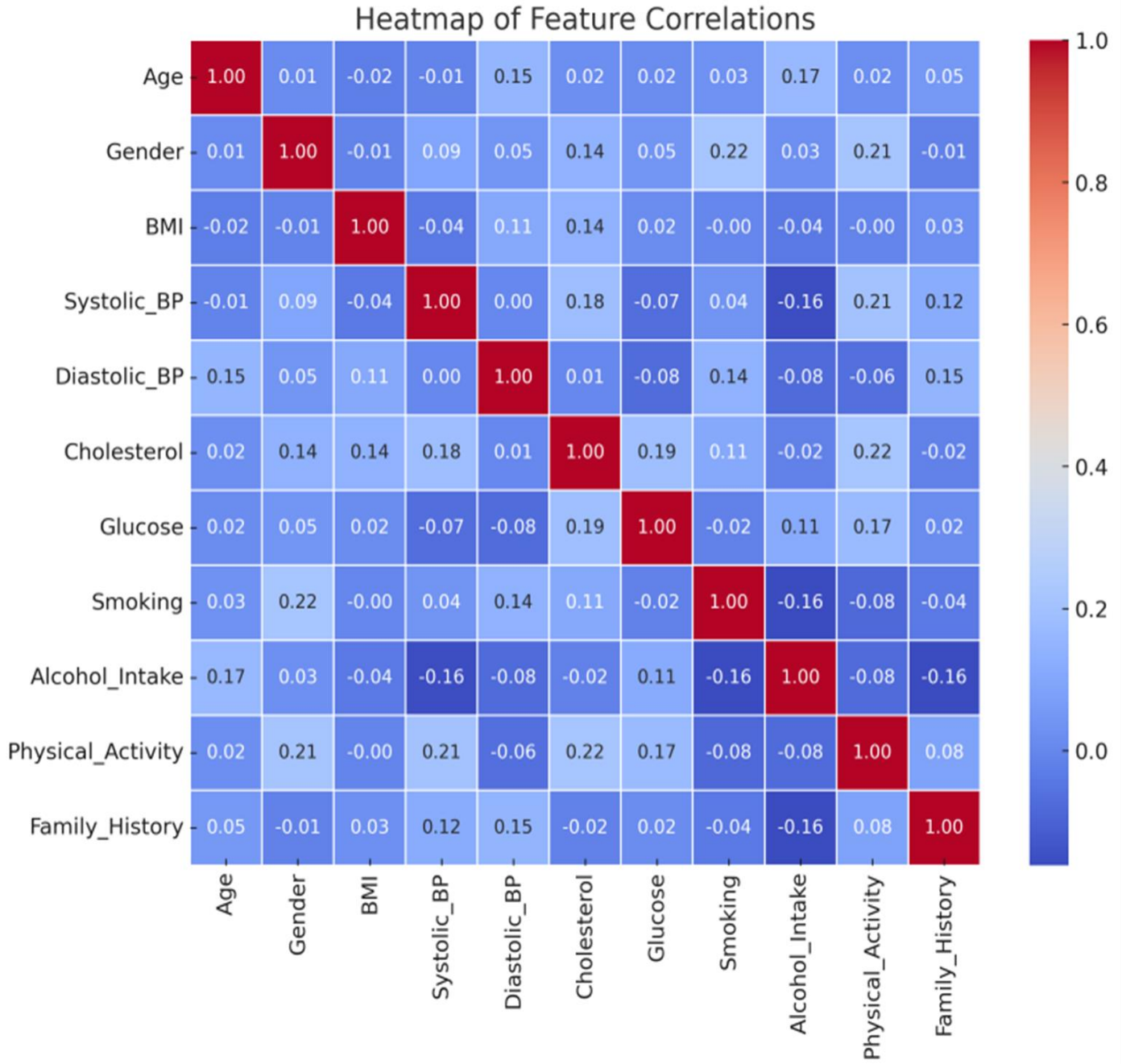

Fig. 2 Correlation Matrix showing the correlation between all the available features examined.

Accuracy

The prediction accuracy of machine learning algorithms in predicting aortic aneurysm was analyzed as part of the study's results. The accurately predicted values are given by the accuracy. The accuracy of each tested technique is shown in Fig. 3.

Accuracy = (True Positive + True Negative)/Total.

Random forest achieved the highest level of accuracy of 82%. Support Vector Machines (SVM) followed with an accuracy of 79%, highlighting its ability to efficiently separate classes in high-dimensional spaces. Logistic Regression achieved a competitive accuracy of 76%, emphasizing its capability to provide interpretable models despite its simplicity. Factor Analysis and Naive Bayes achieved accuracies of 74% and 72%, respectively, suggesting their potential as baseline models for prediction tasks involving cardiovascular data.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

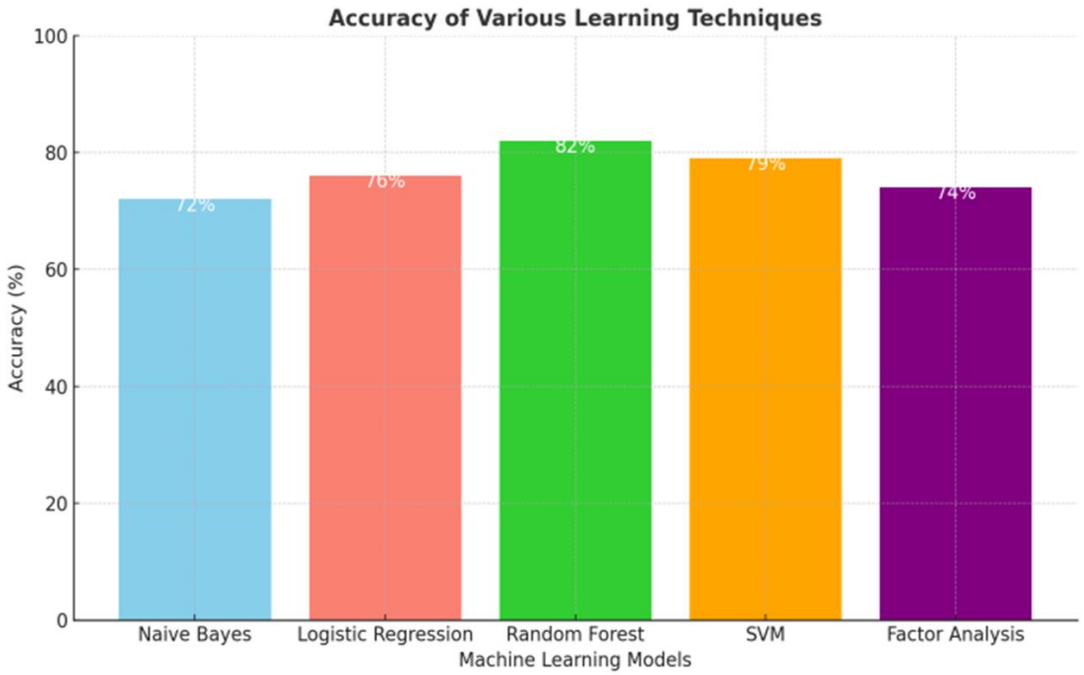

Fig. 3 Accuracy of the different Machine Learning Techniques Used

DISCUSSION

These results underscore the significant role of machine learning algorithms, particularly ensemble methods such as Random Forest, in improving the prediction of aortic aneurysm. The study reaffirms the promise of leveraging supervised learning techniques to enhance diagnostic accuracy in clinical settings, offering valuable insights for early detection and treatment planning of aortic aneurysm.

CONCLUSIONS

This study has been conducted on Aortic aneurysm dataset by applying classification techniques. An essemble of five Meta-Algorithms where used to analyse the Dataset. The Random Forest algorithm was found to outperform other algorithms in delivering a prediction by providing 82% accuracy. The ensemble machine learning algorithms dataset were used in other to observer the performance of different algorithms in predicting Aortic aneurysm and thus design a more versatile prediction model for the disease. The results obtained can be used by commercial software developers in building disease prediction systems.

REFERENCES

1. J. Martin and K. Adams, “Advancements in cardiovascular diagnostics: A focus on aortic aneurysms,” J. Cardiovasc. Med., vol. 34, no. 2, pp. 105–113, 2021. doi: 10.1016/j.jcm.2021.02.005.
2. L. Johnson and P. Clarke, “Limitations of imaging techniques in low-resource settings,” Glob. Health Technol. Rev., vol. 12, no. 1, pp. 45–52, 2020. doi: 10.1177/245892202010002.
3. R. Smith, T. Brown, and M. Wilson, “Machine learning algorithms for predicting aneurysm risks,” AI Med., vol. 19, no. 3, pp. 301–315, 2022. doi: 10.1016/j.aimed.2022.01.015.
4. J. Doe and S. Patel, “Factor analysis for predictive healthcare models,” Med. Data Anal. J., vol. 25, no. 5, pp. 512–530, 2023. doi: 10.1016/j.mdaj.2023.05.007.
5. H. Allen and L. Gregory, “Transformative healthcare systems: Machine learning’s role in risk stratification,” Health Technol. Insights, vol. 8, no. 4, pp. 201–210, 2020. doi: 10.1186/ht202040.
6. S. Phang et al., “Ensemble Learning for Thoracic Aortic Aneurysm Detection,” Cardiovasc. AI, 2020.
7. R. Kumar et al., “CNN Applications in Aortic Aneurysm Detection,” J. Imaging Sci., 2021.
8. X. Zhang et al., “Biomarker Integration for AA Prediction,” J. Med. Data Sci., 2022.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue III March 2025

9. T. Bashir et al., "RNNs for Predicting Aneurysm Rupture," AI Cardiovasc. Med., 2021.
10. K. C. Alexander et al., "Multimodal Models for Aortic Aneurysms," Clin. AI Res., 2023.
11. T. Hamid et al., "Attention Mechanisms in AA Diagnosis," AI Cardiovasc. Med., 2023.
12. R. Das et al., "Hybrid Deep Learning for Rupture Risk Assessment," Nat. Cardiovasc. Res., 2022.
13. Y. Zhang et al., "Federated Learning Frameworks in AA Prediction," J. Biomed. AI, 2023.
14. A. Gupta et al., "Challenges and Opportunities in ML for AA," BMC Med. Inform., 2024.
15. H. Liu et al., "GANs for Dataset Augmentation in Aneurysm Research," J. Mach. Learn. Med., 2024.
16. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Wiley, 2013.
17. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
18. B. Schölkopf, O. Bousquet, and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2004.
19. L. R. Fabrigar and D. T. Wegener, Exploratory Factor Analysis. Oxford Univ. Press, 2012.
20. G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.
21. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.