Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

The Integration of Machine Learning in information technologies: Future Trends and predictions

Mahfuz Alam[1]✉, Md Rafiqul Islam[2], Mir Mohtasam Hossain Sizan[3], and Al Amin Akash[4]

[1]Department of Business Administration, MBA in Business Analytics, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010
[2] Department of Business Administration, MBA in Business Analytics, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010
[3] Masters of Science in Business Analytics, University of North Texas
[4] Bachelor in Computer Science, La Roche University, 9000 Babcock Boulevard Pittsburgh, PA 15202

Corresponding Author: Mahfuz Alam, **E-mail**: jnmahfuz@gmail.com

| ABSTRACT

Hypertension and high cholesterol can cause heart attack and heart failure. Therefore, preventative measures to effectively anticipate, diagnose, and control high hypertension and cholesterol levels are needed to reduce myocardial infarction risk and offer more effective treatment alternatives. The study used new machine learning models (SVM, LR, RF, and NB) to predict cardiac illnesses, which were not used in a prior study. The study utilized four new machine learning models (SVM, LR, RF, and NB) to predict cardiac diseases, which had not been previously studied. The study used a 1970–2023 dataset of UK males with cardiac issues. Results showed that machine learning methods have gained popularity and can predict cholesterol and hypertension. The study shows that machine learning algorithms lower hypertension and cholesterol. Machine learning techniques must be improved and tested on larger datasets.

| KEYWORDS

Machine learning techniques, hypertension, heart diseases, cholesterol, larger dataset, forecasting models

| ARTICLE INFORMATION

ACCEPTED: 20 October 2024 **PUBLISHED:** 10 November 2024 **DOI:** 10.32996/jcsts.2024.6.5.7

I. Introduction

Hypertensive heart disease refers to heart problems that occur because of high hypertension that is present over a long time. Hypertensive heart disease constitutes functional and structural dysfunction and pathogenesis occurring primarily in the left ventricle, the left atrium and the coronary arteries due to chronic uncontrolled hypertension (Masenga and Kirabo, 2023). Cholesterol plays a detrimental role in the pathogenesis of atherosclerosis and cardiovascular disease CVD (Avci et al., 2018). In the medical industry, machine learning techniques are going to be important day by day. The use of machine learning techniques is very powerful to analyze, identify, and anticipate a wide range of diseases. Machine learning is a fascinating field within artificial intelligence that involves training computer machines using vast amounts of data. To enhance the level of patient care and effectively handle extensive volumes of medical data, machine learning algorithms play a vital role (Alarsan and Younes, 2019). Pal et al. (2023) performed a thorough analysis of five different supervised machine-learning algorithms to predict heart disease. Prior research did not predict heart illnesses using SVM, LR, RF, or NB machine learning models. To overcome this limitation, the study uses a larger and more diverse dataset to increase results significance. A robust framework incorporating contemporary machine learning and boosting ensemble approaches improves forecast accuracy. The study has applied SVM, LR, RF, DT, and NB synergies for early detection and risk assessment for prevention and therapy. Improving heart disease prediction and mitigation is a public health priority by using machine learning models SVM, LR, RF, DT, and NB to predict heart health. These approaches predict cardiovascular disease best and provide important information, and Gradient Boost for two years, regression predicts better. The second-best machine learning forecasting method is Random Forecast, and It gives the most accurate five-year hypertension projections. The findings show that the research may improve patient outcomes and cut cardiovascular disease costs, which will considerably benefit healthcare output.

2. Literature Review

The smooth support vector machine model (SSVM) that Wang et al. (2023) created was based on a novel smooth function, and it demonstrated improved classification capabilities. With a prediction accuracy of 91.67%, the SVM model was found to have the highest accuracy of any model when it came to predicting cardiac diseases, as stated by Boukhatem et al. (2022). According to Yong et al. (2023), the GB-based model had superior performance in comparison to other models following the incorporation of feature engineering strategies. Moore and Bell (2022) found that XGBoost performed better than logistic regression when it came to identifying persons who had subsequently experienced a myocardial infarction on other occasions. In their study published in 2022, Özhan and Kuçükakçali discovered that the utilization of the XGBoost model yielded the most encouraging outcomes for heart attacks. When it comes to predicting heart disease, Pratyushaa and Kanimozhib (2022) state that the decision tree algorithm is superior than the closest neighbor approach in terms of performance. When it comes to the prediction of cardiac disease, the smote-Xgboost algorithm approach developed by Yang and Guan (2022) offers substantial benefits. When compared to other methods of machine learning, Shaw and Patidar (2022) state that support vector machines (SVM) have an accuracy of 92.67% when it comes to predicting cardiac disorders. Teja and Veeramani (2022) discovered that support vector machines (SVM) had a significantly lower random forest, logistic regression, and decision tree. According to Theerthagiri and Vidya (2022), the gradient boosting algorithm will function as a much more effective and prominent method for the prediction of cardiovascular ailments inside the future. The decision tree and the support vector machine (SVM) both have their own unique features; the model provides categorized outcomes in order to forecast data (Vijaya et al., 2022). For the purpose of this investigation, the LSTM model was utilized in order to capitalize on the association that exists between the three-dimensional motions of auxiliary points (APs) and points of interest (POI) on the surface of the heart (Zhang et al., 2022). According to Zhou (2022), a low-power heart sound diagnostic processing unit (HSDPU) for wearable auscultation devices that was based on LSTM was able to attain an accuracy of 96.9%.

3. Research Methodology

3.1 Data Collection

The dataset on heart disease encompasses a diverse array of attributes related to lifestyle choices and cardiovascular well-being. The data related to hypertension and cholesterol has been collected from medical history heart disease datasets, which were utilized, collected from the databases of the World Health Organization (WHO) and the British Heart Foundation (BHF). These models have achieved substantial advancements in the accurate forecasting, identification, assessment, and prediction of many medical conditions. The research examined three ensemble-learning-based boosting algorithms to predict heart disease. The experiment's use of the algorithms has been explained as listed below. Ensemble learning methods like Random Forest, AdaBoost, and Gradient Boosting are frequently employed in healthcare research for analyzing intricate datasets and deriving valuable insights. Through the application of these techniques, the study seeks to offer valuable tools for constructing and implementing.

3.2 Decision Tree

The DT classifier effectively extracts valuable insights from data using a reliable decision tree algorithm. By employing a series of queries, it can classify the data and uncover valuable distinctions that may not be immediately apparent. When it comes to predicting the value of a target variable, the key is to train a model that can grasp the underlying decision rules derived from the properties of the data.

3.3 Random Forest

Researchers commonly employ the mean of many individuals' decisions when making a selection. Methods such as these can be classified as ensemble methods. The Random Forest classifier generates a novel model by aggregating the best or most often occurring predictions from Decision Trees (DTs). Its precise outcomes render it uncomplicated but robust. The ensemble approach employed in this study utilizes DT forests. The Random Forest classifier is a crucial component for both regression and classification tasks after it has been established.

3.4 Structural Vector Model

Support vectors involve the utilization of a hyperplane to separate datasets into two distinct groups. Through the assistance of vectors, which are the farthest points in the categories, the identification of the optimal hyperplane becomes more convenient. When dealing with non-linearly separable data, a high-dimensional hyperplane becomes necessary.

3.5 Gradient Boost Regressor

Ensemble machine learning uses gradient boosting, and this technique creates a more reliable prognostic model using Decision Trees and weak learners. The gradients' mathematical formulations rely on the loss function. Regression tasks in machine learning employ mean squared error, while classification uses cross-entropy. To minimize loss, the model improves its predictions as it investigates the issue, revealing intricacies. Table 1. Description of variables, codes and units of measurement.

Table 1. Classification of variables, codes, and units

Variable	Code	Unit
Age-standardized prevalence of hypertension	ASPH	mmHg
Mean total cholesterol	MTC	mmol/L

Age-normalized hypertension prevalence permits fair comparison of groups of different ages. It shows population age differences when connecting, and the most prevalent method is direct normalization. The approach involves comparing age group prevalence rates to the population's age distribution.

4. Results and Discussions

4.1 Descriptive Statistics

Table 2 represents the different metrics used to summarize and illustrate a dataset's attributes. The mean represents the average level of cholesterol in the sample. When arranged in ascending order, the middle determined the dataset's central value.

Table 2. Descriptive statistics

Variable	Mean	Std	Min	Q1	Median	Q3	Max
ASPH	0.22646	0.04309	0.23247	0.2124	0.26346	0.30522	0.31422
MTC	4.27228	0.37534	2.94234	4.23456	4.34275	4.48632	4.82422

The deviation represents the dispersion of the hypertension and cholesterol levels around the mean. The normal of the squared contrasts from the mean is shown as changing. Normal percentiles include the 22th, 50th (middle), and 75th percentiles. These measurements have provided a comprehensive overview of the dataset, including the distribution, dispersion, and focal inclination of the levels of hypertension and cholesterol. Machine Learning models have demonstrated strong performance in the identification of undetected hypertension and its associated variables within the South Asian region (Siddiquee, 2023). The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have been used as evaluation criteria for the comparison of these regression models (Ameer et al., 2019). Table 3 summarizes the model performance results for age-standardized prevalence hypertension.

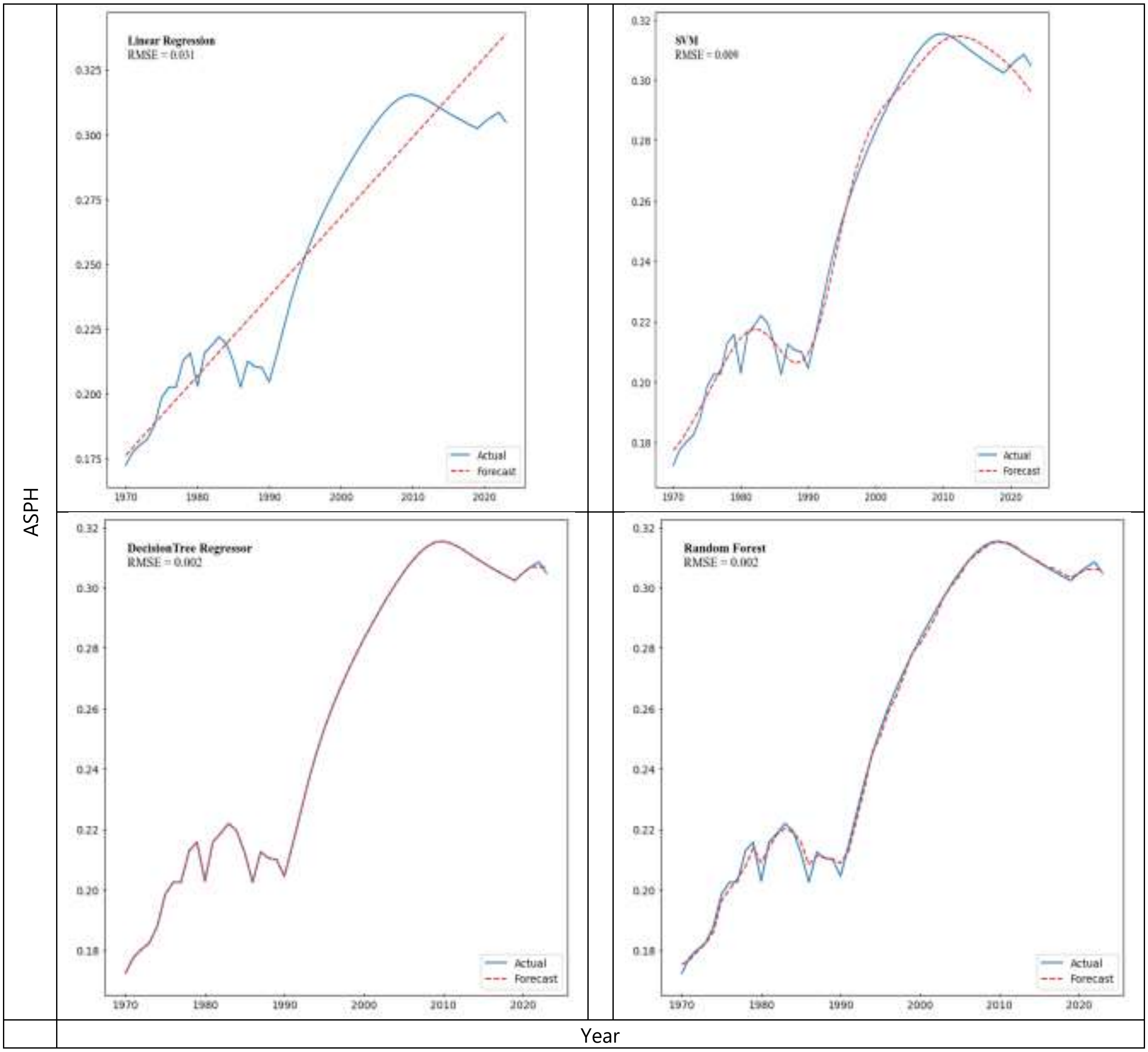
Table 3. Assessment of ASPH forecasting models' performance metrics

Model	2-Years forecast			5-Years forecast		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Linear Regression	0.03094	0.03075	0.10034	0.03444	0.03422	0.11202
Decision Tree	**0.00189**	**0.00188**	**0.00612**	0.00273	0.00231	0.00752
Gradient Boost Regressor	**0.00188**	**0.00188**	**0.0061**	0.00267	0.00224	0.00732
Random Forecast	0.00196	0.00188	0.00611	**0.0021**	**0.00182**	**0.00596**
SVM	0.00888	0.00887	0.02893	0.02162	0.02014	0.06578

Note: Bold values reflect the method's most accurate outcomes.

The results reveal that the decision Tree is the second machine learning tool for best forecasting results for the next 2 years with minimum values of RMSE, MAE, and MAPE metrics. Random Forecast shows hypertension prediction best performance for the next 5 years with minimum values of RMSE, MAE, and MAPE metric results. Random Forecast is the second-best machine learning forecasting technique which performed the best analysis of hypertension for the next 5 years forecast with minimum values of RMSE, MAE, and MAPE metric results. Figure 1 represents the comparison of the time series of estimated and observed age-standardized prevalence hypertension using different machine learning models.

Figure 1. Comparison machine learning model-estimated prevalent hypertension time series



Note: The prevalent hypertension time series using different machine learning methods.

The use of machine learning techniques has proven to be highly beneficial in the prevention and management of hypertension in Africa, Southeast Asia, and low-income nations (Boateng & Ampofo, 2023). The scatter plots of RMSE values 0.0031, 0.004, 0.002, and 0.009 for Linear Regression, Random Forest, and SVM, respectively, indicate the accuracy of these machine-learning techniques in predicting the age-standardized prevalence of hypertension. These values suggest that Random Forest performs very well in estimating hypertension prevalence, with Random Forest being the most accurate among them. SVM, while still accurate, is slightly less so than the other techniques. In previous years, these techniques have been effective in estimating and observing the age-standardized prevalence of hypertension compared to other machine-learning techniques. Figure 2 represents the scatter plots of observed and estimated age-standardized prevalence hypertension using different machine learning models.

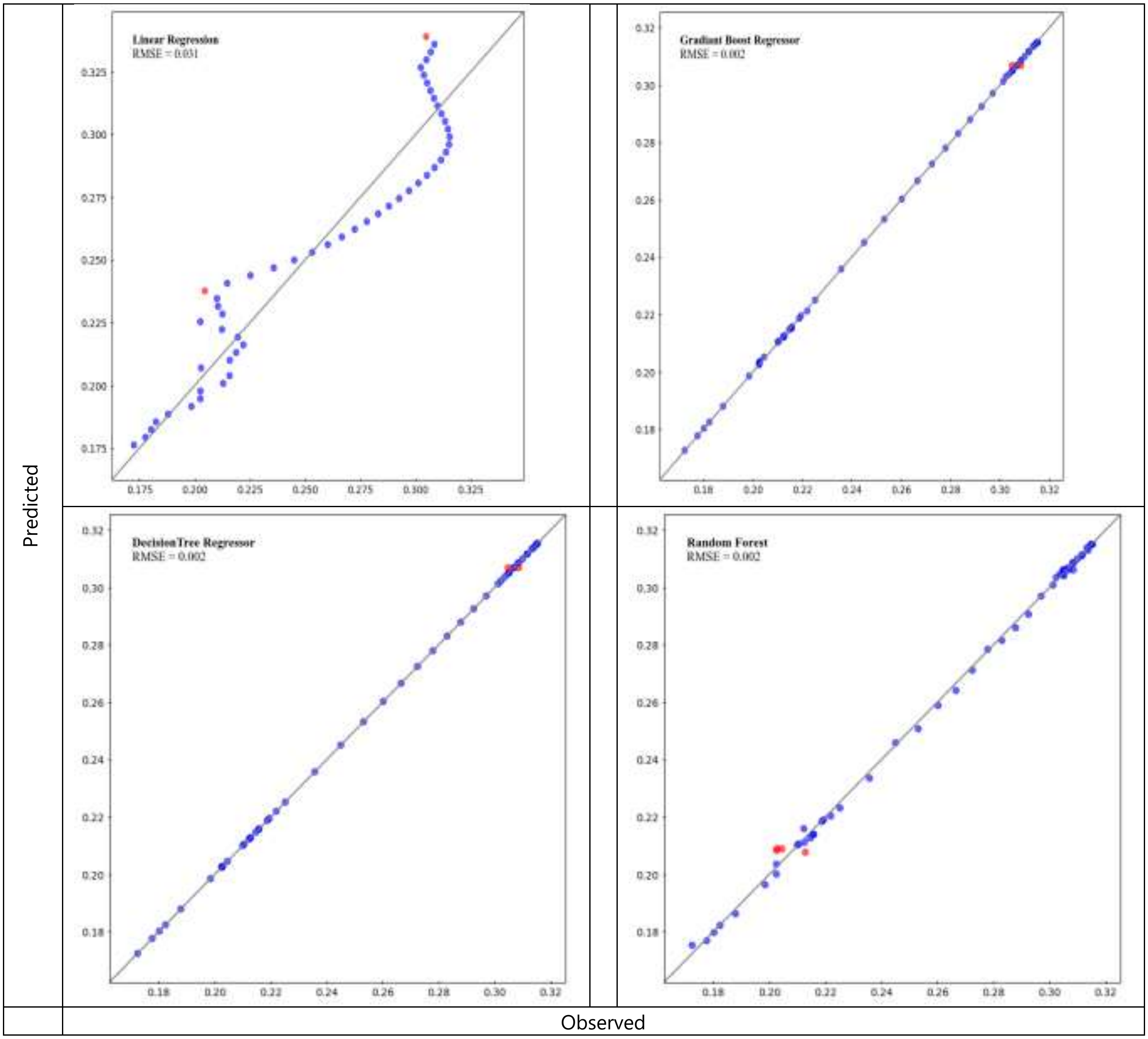


Figure 2. Scatter graphs of observed hypertension using machine learning algorithms

The scatter plots observation of RMSE values 0.0031, 0.004, 0.002, and 0.009 for the Linear Regression, Random Forest, and SVM models, respectively, indicate that machine learning techniques can effectively predict the occurrence of age-normalized hypertension. The Random Forest models have demonstrated efficacy in determining the incidence of hypertension. However, the Random Forest model has the highest level of reliability. SVM is precise, yet less precise than alternative methodologies. The efficacy of these approaches in evaluating and identifying the age-normalized prevalence of hypertension in younger individuals has been regularly demonstrated, hence confirming their suitability for tackling challenging tasks.

4.2 Forecasting of Mean Total Cholesterol

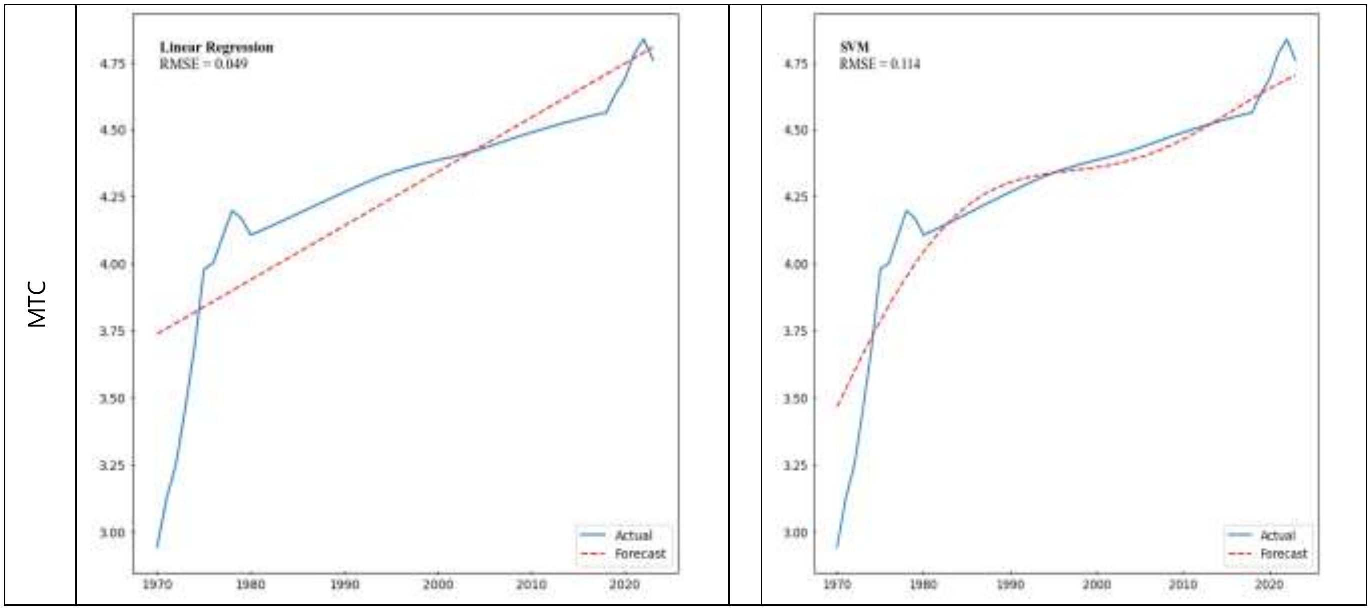
Table 4 represents the mean total cholesterol data set performance metrics of different forecasting models. Exhibition measurements for estimating models in the mean complete cholesterol informational collection typically include metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) values. These measurements are used to evaluate the accuracy and effectiveness of the models in predicting the average total cholesterol values. Table 4 represents the lists of the best hypertension and cholesterol predicting methods for the next 2 and 5 years.

Table 4. Effectiveness of prediction models on mean total cholesterol

Model	**2-Years forecast**			**5-Years forecast**		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Linear Regression	0.0433	0.04327	0.01027	**0.06467**	**0.05519**	**0.01233**
Decision Tree	**0.04264**	**0.03918**	**0.00814**	**0.19099**	**0.237**	**0.03712**
Gradient Boost Regressor	0.04279	0.03918	0.00814	0.1913	0.23734	0.03719
Random Forecast	**0.04234**	**0.03923**	**0.00812**	0.19537	0.18232	0.02411
SVM	0.1139	0.10377	0.02155	0.21005	0.19868	0.0423

Note: Bold values reflect the method's most accurate results

Linear regression, decision tree, and gradient boosting fared better for 2-year projections of mean total cholesterol levels, with reduced RMSE, MAE, and MAPE values. The SVM and random forecast models performed poorly for short- and long-term forecasts. Linear regression made the best 5-year predictions, whereas Random Forecast made the best short-term predictions. Random projection has the best mean total cholesterol projection for the next two years with the least RMSE, MAE, and MAPE values. Decision Tree, another powerful machine learning method, predicts mean total cholesterol for the following two years. With the lowest RMSE, MAE, and MAPE values, Linear Regression forecasts mean total cholesterol for the next five years best. While Decision Tree is another powerful machine learning approach for predicting 5-year mean total cholesterol values. Figure 3 shows scatter plots of observed and predicted mean total cholesterol using machine learning algorithms.



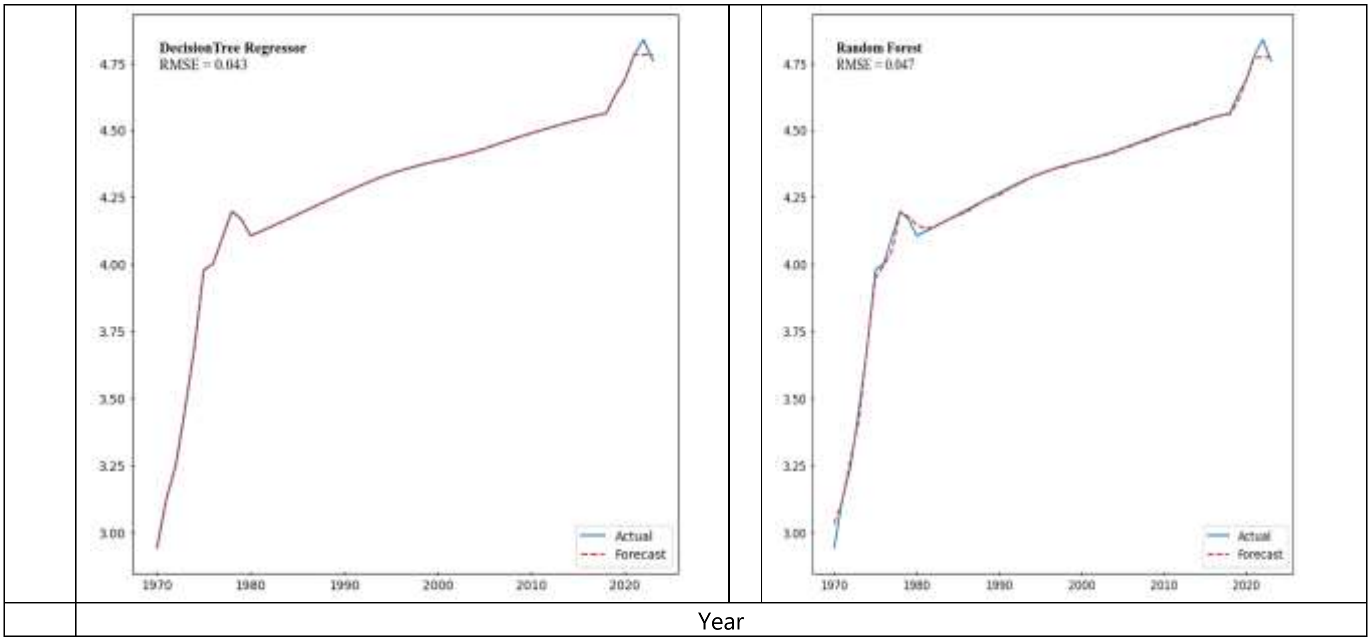


Figure 3. Machine learning model comparison of estimated cholesterol time series

The plots reveal that the Random Forest predicted cholesterol levels best of the three machine learning algorithms tested. Linear Regression performed well. Although the AdaBoost Regressor somewhat deviated from the curve, Linear Regression showed a non-linear connection between cholesterol and the dependent variable. Due to its low root-mean-squared error (RMSE) of 0.047, the Random Forest technique outperformed other methods. Linear Regression and LSTM both performed well, but Random Forest estimated cholesterol prevalence more accurately. SVM RMSE has increased in recent decades. Table 5 represents the future sight: precise forecasts for the years ahead for both hypertension and cholesterol.

Table 5. Future Sight: Precise Forecasts for the Years Ahead.

Variable	2-Years ahead forecast	5-Year ahead forecast
ASPH	0.30502	0.30524
MTC	4.55808	4.92348

Modelling future patterns in hypertension and cholesterol levels entails gathering and examining past data through the application of machine learning methodologies. These transdisciplinary machine learning algorithms can forecast hypertension and cholesterol levels for the next two and five years. The findings in Table 6 demonstrate the verification of the machine learning models, confirming their precision. Gradient Boost Regressor performs best to forecast ASPH for next two years, and it has performed as best forecasting technique for the prediction of ASPH for next five years. Random Forecast performs best to forecast MTC for next two years, and Linear Regression has performed as best forecasting technique for the prediction of MTC for next five years. Scatter plots of observed and estimated age-standardized prevalent hypertension using machine learning algorithms. Figure 4 represents the Scatter plots of observed and estimated age-standardized prevalent hypertension using machine learning algorithms.

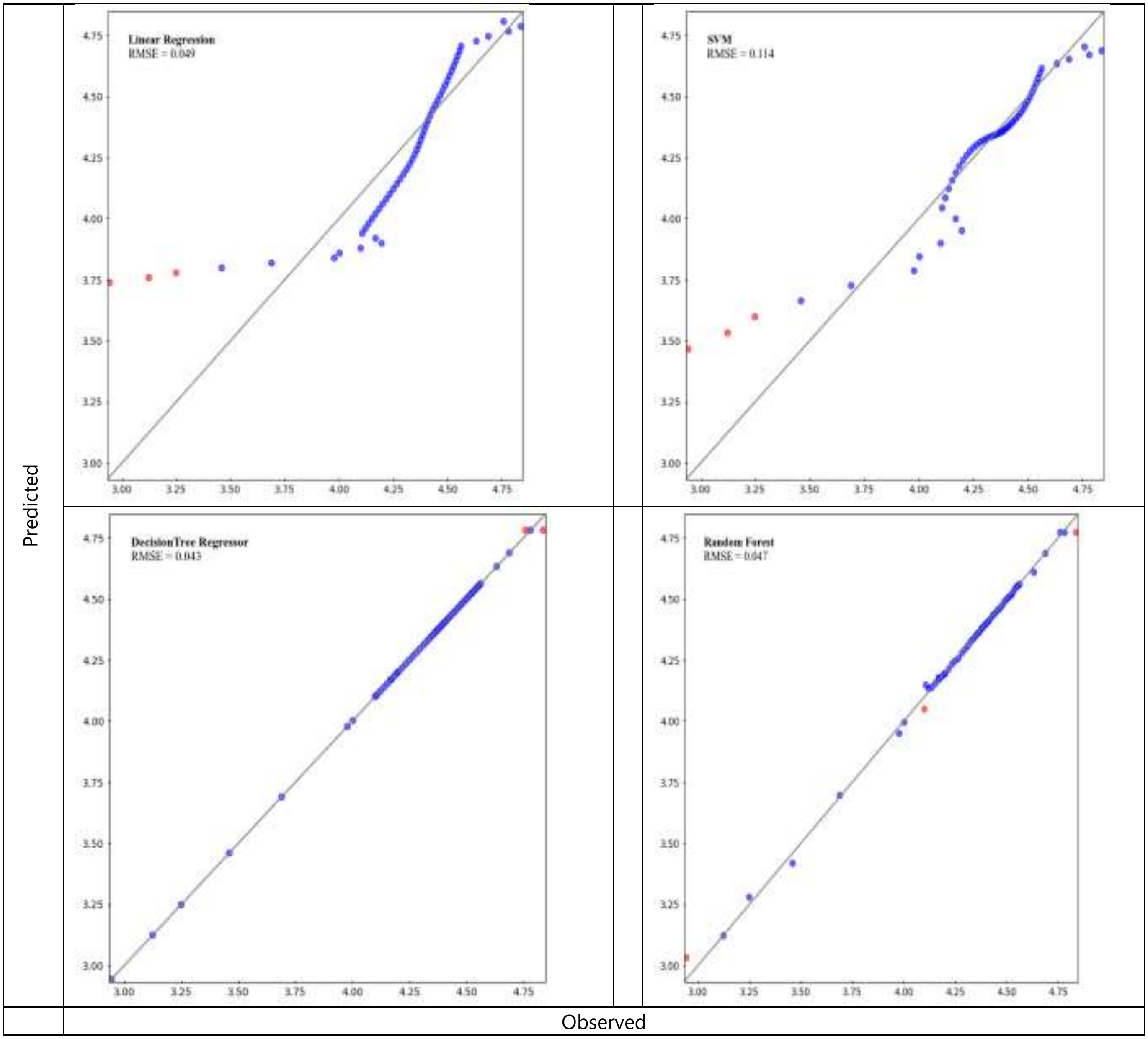


Figure 4. Various machine learning algorithms predicted and observed cholesterol

4.3 Discussions

The findings indicate that the nonlinear machine learning models demonstrated superior performance compared to the baseline linear regression for hypertension future prediction. The findings underscore the efficacy of employing ensemble and nonlinear modelling methodologies in addressing this particular issue. Random Forecast is a machine learning forecasting approach that ranks second in terms of performance. It provides the most accurate analysis of hypertension forecasts for the next 5 years, with the lowest values of RMSE, MAE, and MAPE metrics. For 2-year predictions, the findings indicate that linear regression, decision tree, and gradient boosting, models outperformed other methods in predicting mean total cholesterol levels. In both short- and long-term forecasts, the SVM, and random forecast models exhibited worse performance. Linear regression yielded the highest level of accuracy in 5-year projections of cholesterol, but Random Forecast demonstrated ideal performance in short-term forecasting scenarios. The Random Forecast demonstrates the highest level of performance in predicting the mean total cholesterol for the next two years, as indicated by the lowest values of RMSE, MAE, and MAPE metrics. The Decision Tree algorithm is a second machine learning technique that has demonstrated excellent efficacy in predicting the mean total cholesterol values for two years of cholesterol prediction. The Linear Regression has the highest level of performance in forecasting mean total cholesterol over the next five years, as seen by the lowest values of RMSE, MAE, and MAPE metrics.

5. Conclusion

A complete investigation of ensemble machine learning approaches, notably boosting algorithms, for early prediction of two heart diseases is provided by the study. The dataset was cleaned up using advanced methods to remove outliers and damaged data. Hypertension and cholesterol were predicted using nine machine learning algorithms and comparable methods. These algorithms' efficiency was assessed using statistical markers. The study found that machine learning methods can predict cholesterol and hypertension in the future. This study examines new ML models for heart disease prediction. For hypertension prediction, nonlinear machine learning models outperformed baseline linear regression. The findings demonstrate the effectiveness of ensemble and nonlinear models in tackling the issue. Forecasting for two years is better using the Gradient Boost Regressor. The decision Tree algorithm is the second machine learning method chosen for excellent two-year prediction. Random Forecast is a second-best machine learning forecasting method. It predicts hypertension for the following five years most accurately. Linear regression, decision tree, gradient boosting, AdaBoost, and XGBoost models predicted mean total cholesterol levels better over 2 years. SVM and random forecast models performed poorer in short- and long-term forecasts. In 5-year cholesterol forecasts, linear regression was most accurate, whereas Random Forecast performed best in short-term forecasting. The Random Forecast predicts mean total cholesterol for two years best. Another machine learning method that has accurately predicted mean total cholesterol readings for two years is the Decision Tree algorithm. Linear Regression predicts five-year mean total cholesterol best. Another machine learning method that accurately predicts 5-year mean total cholesterol readings is the Decision Tree technique.

5.1 Future Research Suggestions

The study suggests machine learning models might enhance diabetes and blood sugar prediction and therapy. Complete dataset surveys and enhancements are needed, and the potential to run these models as software on mobile phones or diabetes care devices is great. The treatment works for diabetes and associated disorders.

References

[1] Avci, E., Dolapoglu, A., & Akgun, D. E. (2018). Role of cholesterol as a risk factor in cardiovascular diseases. Cholesterol-Good, Bad and the Heart, 10.

[2] Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. Journal of big data, 6(1), 1-15.

[3] Boukhatem, C., Youssef, H. Y., & Nassif, A. B. (2022, February). Heart disease prediction using machine learning. In 2022 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-6). IEEE.

[4] Moore, A., & Bell, M. (2022). XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. Clinical Medicine Insights: Cardiology, 16, 1-6.

[5] Masenga, S. K., & Kirabo, A. (2023). Hypertensive heart disease: risk factors, complications and mechanisms. Frontiers in Cardiovascular Medicine, 10, 1204275.

[6] Özhan, O., & Küçükakçali, Z. (2022). Estimation of risk factors related to heart attack with XGBoost that machine learning model. Middle Black Sea Journal of Health Science, 8(4), 582-591.

[7] Pal, M., Parija, S., & Mohapatra, R. K. (2023). Heart Disease Risk Prediction Using Supervised Machine Learning Algorithms. In Microelectronics, Circuits and Systems: Select Proceedings of Micro2021 (pp. 122-134). Singapore: Springer Nature Singapore.

[8] Pratyushaa, M., & Kanimozhib, K. V. (2022). Heart Disease Prediction Using Decision Tree in Comparison with k-Nearest Neighbor to Improve Accuracy. Advances in Parallel Computing Algorithms, Tools and Paradigms, 41, 231-236.

[9] Shaw, S. K., & Patidar, S. (2022). Heart disease diagnosis using machine learning classification techniques. In Inventive Communication and Computational Technologies: Proceedings of ICICCT 2022 (pp. 445-460). Singapore: Springer Nature Singapore.

[10] Teja, P. P. S., & Veeramani, T. (2022). Comparing the Efficiency of Heart Disease Prediction using Novel Random Forest, Logistic Regression and Decision Tree and SVM Algorithms. Cardiometry, (22), 1431-1439.

[11] Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. Expert Systems, 39(9), e13064.

[12] Vijaya Saraswathi, R., Gajavelly, K., Kousar Nikath, A., Vasavi, R., & Reddy Anumasula, R. (2022, February). Heart Disease Prediction Using Decision Tree and SVM. In Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021 (pp. 69-78). Singapore: Springer Nature Singapore.

[13] Wang, J., He, F., & Sun, S. (2023). Construction of a new smooth support vector machine model and its application in heart disease diagnosis. Plos one, 18(2), 1-14.

[14] Wang, Y. (2023, March). Risk assessment for heart failure using multiple linear regression model. In Second International Conference on Biological Engineering and Medical Science (ICBioMed 2022) (Vol. 12611, pp. 1228-1235). SPIE.

[15] Yang, J., & Guan, J. (2022). A heart disease prediction model based on feature optimization and smote-Xgboost algorithm. Information, 13(10), 1-15.

[16] Yong charoenchaiyasit, K., Arwatchananukul, S., Temdee, P., & Prasad, R. (2023). Gradient Boosting Based Model for Elderly

Heart Failure, Aortic Stenosis, and Dementia Classification. IEEE Access.

[17] Zhang, W., Yao, G., Yang, B., Zheng, W., & Liu, C. (2022). Motion prediction of beating heart using spatio-temporal LSTM. IEEE Signal Processing Letters, 29, 787-791.

[18] Zhou, J., Lee, S., Liu, Y., Liu, T., Tse, G., & Zhang, Q. (2021). Predicting Stroke and Mortality in Mitral Regurgitation: A Gradient Boosting Approach. medRxiv, 1-26.

[19] Zhou, W., Wang, A., & Yu, L. (2022, September). A heart sound diagnosis processing unit based on LSTM neural network. In 2022 IEEE 4th International Conference on Circuits and Systems (ICCS) (pp. 210-215). IEEE.