UNIZIK JOURNAL OF ENGINEERING
AND APPLIED SCIENCES
(UJEAS)

Research Article

Prediction of Heart Disease using Autoencoder with LightGMB and Gradient Boosting

Gbenga O. Ogunsanwo, Olalere A. Abass and Emmanuel. O. Taiwo.

Special Issue
A Themed Issue in Honour of Professor Onukwuli Okechukwu Dominic (FAS).

This special issue is dedicated to Professor Onukwuli Okechukwu Dominic (FAS), marking his retirement and celebrating a remarkable career. His legacy of exemplary scholarship, mentorship, and commitment to advancing knowledge is commemorated in this collection of works.

Edited by
Chinonso Hubert Achebe PhD.
Christian Emeka Okafor PhD.

UJEAS

UNIZIK JOURNAL OF ENGINEERING
AND
APPLIED SCIENCES

UNIZIK Journal of Engineering and Applied Sciences 4(1), March (2025), 1619-1630
Journal homepage: https://journals.unizik.edu.ng/index.php/ujeas
PRINT ISSN: 2992-4383 || ONLINE ISSN: 2992-4391

Prediction of Heart Disease using Autoencoder with LightGMB and Gradient Boosting

Gbenga O. Ogunsanwo[1*], Olalere A. Abass[2] and Emmanuel. O. Taiwo[3]
[1,3]Department of Computer Science, Tai Solarin University of Education, Ijebu-Ode, Ogun State, Nigeria.
[2]Department of Computer Science, Sikiru Adetona College of Education, Science & Technology, Omu-Ajose, Ogun State, Nigeria.
[*]Corresponding Author's E-mail: ogunsanwogo@tasued.edu.ng

Abstract
Cardiovascular disease (CVD) refers to heart disease. CVD is seen to cause majority of death in world. Because of this problem many researchers have been drawn to this area to develop models and systems to predict occurrence of heart disease for early treatment. This study developed a Heart disease predicative model using Autoencoder with LightGBM and Gradient Boosting. The Dataset used was gotten from kaggle.com. One Hot Encoding, SMOTE were used to pre-processed the dataset. Feature extraction was done using Autoencoder. Two classification methods: LightGBM and Gradient Boosting were employed to build the predictive model. The result shows that Autoencoder perform very well with low values of MSE, RMSE, MAE and High Value of F2 score. The result of LightGBM shows a specificity of 95.7%, precision value of 95.7%, recall value of 94.7% ,F1 value of 95.2% , AUC value of 0.99 and Accuracy value of 95 While the results of Gradient Boosting shows Specificity of 91.7%, Precision value of 91.5%, Recall value of 90.0% ,F1 value of 91.0% , AUC value of 0.96 and Accuracy value of 90. The study concluded that LightGBM perform better that Gradient Boosting. The Model is recommended to the health sector management to guide their decision making. Its potential integration with predictive model and clinical validation will assist greatly in improving the heart disease diagnosis and prevention. Further research could be done with more validating metrics, more deep learning techniques.

Keywords: Cardiovascular *Disease,* Autoencoder**,** LightGBM, Gradient Boosting

1. Introduction

Heart is an essential muscle responsible for circulating blood throughout the body. Heart disease (HD) is also refers to as cardiovascular disease (CVD). World Health Organization. (2021) reported that CVD. Is one the leading cause mortality globally, causing a lot of stress on the healthcare system and economies. It is very crucial to the cardiovascular system. HD ranks as a major cause of death in the world. As reported by World Health Organization, HD and stroke lead to 17.5 million deaths each year worldwide. Over 75% of these fatalities occur in countries with middle- and low-income populations. Additionally, heart attacks and strokes make up 80% of all deaths attributed to cardiovascular diseases (CVDs). It is important to say that early and accurate diagnosis of HD is important for timely intervention in order to increase the patient survival rate. Moreover, the method that depend on subjective clinical judgments and extensive testing is known as traditional diagnostic which comes with many flaws so also its expensive and time consuming. In recent times, machine learning (ML) has shown to be a bright option for developing predictive models for numerous healthcare applications in which HD is not exempted (Abass et al, 2020; Ogunsanwo 2024).

ML algorithms are good in learning complex patterns and relationships among variables from voluminous dataset that position them to identify people at high risk of developing CVD based on different risk factors and clinical features. Although several studies have used the potential of ML in HD prediction such as Pradip & Atharva (2024) developed a model to predict HD. The study employed Naive Bayes, Logistic Regression (LR), and SVM to construct models. Cross validation approach was applied to increase the accuracy of the models. The results indicated that SVM model outperformed other two models. Akare et al. (2024) conducted a study to develop heart prediction system. DT and LR algorithms were utilized to build models. LRn model outperformed the other model with the accuracy of 90.71%. Bharani et al. (2024) carried out a study on comparative analysis of ML algorithms for predicting heart disease. The study applied six ML techniques such as KNN, Naive Bayes, RF, SVC , DT and LRn. Experimental results showed that Random Forest performed better than the others models in terms of accuracy. Saha et al. (2024) developed a system to forecast heart disease. The study used dataset (BRFSS 2021) obtained from the Center for Disease Control and Prevention (CDC); it consists of more than 400 thousand instance with 304 features. GridSearchCV with 10-fold cross-validation was applied to identify the best model. The predictive model was created with DT, KNN, RF, LR, and EGB. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to validate models. RF produced best outcome with accuracy of 91% and ROC-AUC of 0.91. Bombale et al. (2024) developed a model to predict HD. The authors utilized SVM, KNN, DT, XGBoost and RF algorithms to build predictive models. The results uncovered that RF model had highest accuracy compared to other models.

Kardam et al. (2024) conducted a study on HD prediction. The study applied various ML techniques to develop models. The dataset used was obtained from Kaggle repository. The study employed evaluation metrics such as the confusion matrix, accuracy, precision, recall, and f1-score to validate the models. The results indicated that extreme gradient boosting classifier model outperformed other models with 81% accuracy. Logabiraman et al. (2024) proposed a system to predict HD. The researchers utilized both ML and hybrid ML techniques to construct predictive models. The dataset was collected from medical records and hospitals. Experimental result revealed that hybrid models achieved highest accuracy compared to ordinary ML models. Divya et al. (2024) conducted a study to forecast HD. The dataset obtained from Kaggle repository contains 303 instances and 8 attributes. Naive Bayes, SVM and DT techniques were employed to build predictive models. The results showed that Naive Bayes model performed better than the other models with an accuracy of 93.5%. However, there is still need for improvement in the field of HD prediction using ML. For instances, some existing model may suffer from over-fitting or bias due to imbalanced datasets. So there is need for development of more robust model to gain insights into the underlying factors contributing to CVD.

This study aims to address these challenges by developing a model for HD prediction using two ML algorithms and a rich dataset from Kaggle. The study will use feature-engineering strategies such as one Hot-encoding SMOTE to correct the bias due to imbalanced in the dataset and feature extraction with Autoenconder in order to improve the performance of the model. The results of the findings will assist in early detection and personalized risks assessment of CVD and serves as a guide to researchers in this field of HD prediction.

2.0 Materials and methods

This section talks more about the detailed explanation of the materials and methodologies used in this study to achieve the research objectives. The flow diagram (FD) of the study is shown in Figure 1. The FD provides a structured approach that guide the research process. The processes include data collection, data preprocessing, model training, validation and data exploratory.

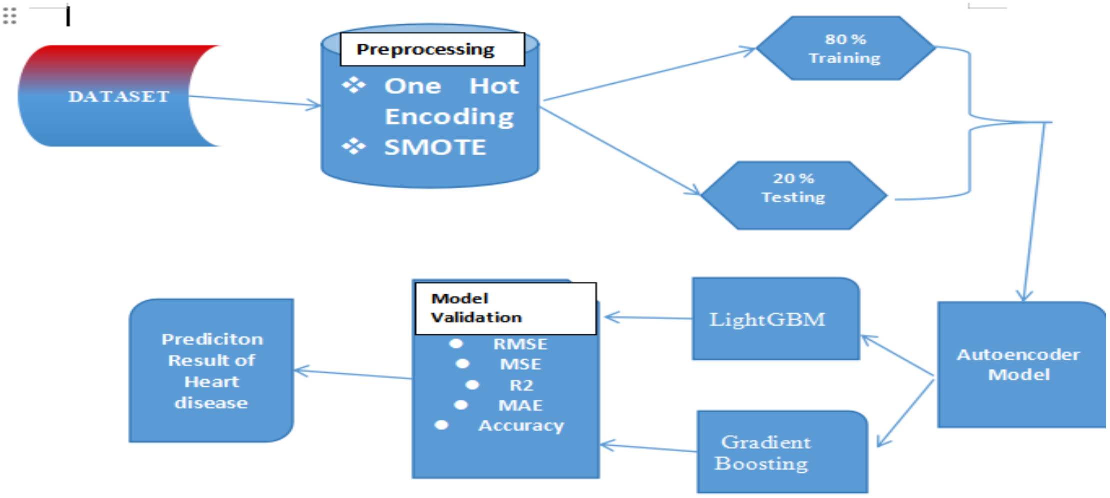


Figure 1: FD of the study

2.1 Data Collection

The Heart dataset used in this study was downloaded from Kaggle repository, https:// www.kaggle.com/dataset. The dataset contains thirteen features and 1888 instances. A description of the dataset is provided in Table 1

Table 1: Description of each features in the dataset used

Parameter	Description	Data Type
Age	Age of the patient	Numerical (Integer)
Sex	Gender of the patient (0 for female, 1 for female)	Categorical (Binary)
Cp	pain in chest types(0,1,2,3)	Categorical (Binary)
Tresbps	Blood pressure due to unrest(in mmHg on admission to the hospital)	Numerical (Integer)
Chol	cholesterol in mg/d	Numerical (Integer)
Fbs	blood sugar > 120mg/d (1 true, 0 false)	Categorical (Binary)
restecg	Resting electrocardiographic results (0,1,2)	Categorical (Ordinal)
thalachh	Highest heart bit achieved	Numerical (Integer)
exang	Exercise caused byangina (1=yes; 0=no)	Categorical(Binary)
oldpeak	ST depression caused by exercise relative to rest	Numerical (Float)
Slope	The gradient of the highest exercise St segment (0,1,2)	Categorical (ordinal)
Ca	Number of prima vessels (0-3) colored by radioscopy	Numerical (Integer)
Thal	Thalassemia (3=normal; 6= fixed defect; 7= reversible defect)	Categorical Nominal)
target	Occurrence of HD presence (0=no, 1 =yes)	Categorical (Binary)

Data Preprocessing

The preprocessing was done on the dataset like encoding categorical features such as sex, cp, fbs, restecg, exang, slope and that was converted into numerical representation for the Autoencoder to process them.

One-Hot Encoding in this the study each category of a categorical feature is transformed into a new binary features (0 or 1) . for instance, the chest pain type (CHP) feature with four categories (0,1,2,3) was transformed into four binary features chp_0, chp_1, chp_2 and chp_3.

SMOTE Synthetic Minority Oversampling Techniques is a well-known oversampling techniques used to solve class imbalance in the HD prediction datasets. When there is more instances of patients without HD compared with patients with HD this phenomenon is known as imbalance in the dataset. This imbalance can affect the performance of the HD predictive model and can lead to biased models (Johnson, & Khoshgoftaar, 2019). The dataset was subjected to SMOTE to correct these issues.

Autoencoders are a example of neural network used for unsupervised learning tasks. They work by learning an efficient representation that is encoding of the input data and used this to reconstruct the original data. Autoencoder is divided into two types namely: Encoder and Decoder. The Encoder consists of different layers of neurons that gradually reducing the dimensionality of the data, it is used to compress the input data into a lower dimensionality of the data. The Decoder also consists of multiple layers of neurons that also gradually reducing the dimensionality of the data, it received the encoded representation from encoder and re construct the original input. (Hinton. & Salakhutdinov ,2006). The Autoencoder was used as feature extraction in this study and the autoencoder was able to reconstruct the image form the original image perfectly as seen in Figure 2

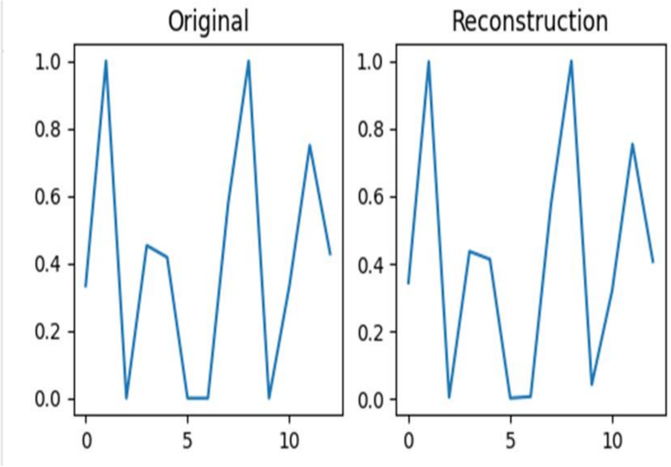


Figure 2 Autoencoder features extraction for Original and reconstruction

Light gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework that uses tree- based learning algorithms. It is designed to be distributed which well suited for large complex tasks. It is known for fast training speed when compared to other gradient boosting. It also makes use of minimum memory space as a result that its uses histogram based algorithms to store data that reduce memory consumption (Meng et al, 2017). LightGBM was used for the classification of the HD in this study and the Mathematical model is seen in Equation 1

$F(x) = \Sigma_i f_i(x)$ (1)

where:

$F(x)$ is the final prediction for input x.

$f_i(x)$ is the prediction of the i-th decision tree.

Σ_i represents the sum over all trees in the ensemble

The Validating Metric Used to validate the Models

Mean Squared Error (MSE) evaluates the squared difference between the original data points and their reconstructions by the autoencoder as seen in Equation 2

$MSE = (1/n) * \Sigma(yi - \hat{y}i)^2$ (2)

Where:

n is the number of data points

yi is the actual value of the i-th data point

ŷi is the predicted (reconstructed) value of the i-th data point

Root Mean Squared Error (RMSE) is the square root of the MSE. It used to measures the magnitude of the errors between original data and the reconstruction as seen Equation 3.

RMSE = √MSE (3)

R-squared (R2) is also known as the coefficient of determination, it represents the proportion of variance in the target variables (original data) which is explained by the autoencoder prediction (reconstructions) as seen in Equation 4

R2 = 1 - (SSres / SStot) (4)

Where:

SSres is the sum of squared residuals that is difference between actual and predicted values
SStot is the total sum of squares that is difference between actual values and the mean of the actual values.

A confusion matrix is a table that is used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions made by the model compared to the actual outcomes;

Accuracy: (TP + TN) / (TP + TN + FP + FN) (5)
Precision: TP / (TP + FP) (6)
Recall (Sensitivity): TP / (TP + FN) (7)
Specificity: TN / (TN + FP) (8)
F1-Score: 2 * (Precision * Recall) / (Precision + Recall) (9)
Where

TP = True Positive
TN= True negative
FP = False Positive
FN False Negative

3.1 Model Training

The HD dataset used was acquired from Kaggle and preprocessed using techniques such as One-Hot Encoding and SMOTE thereafter the dataset was divided into 80% training and 20% testing as shown in Figure 3.

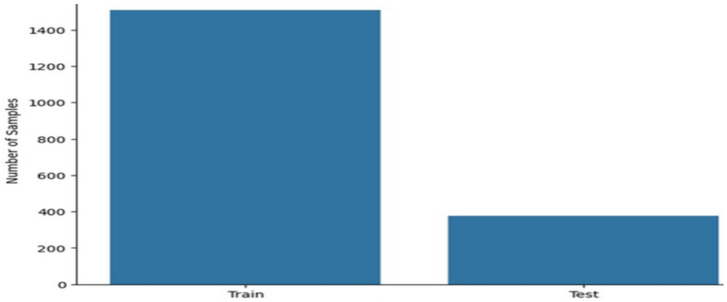


Figure 3 Dataset divided into training and Testing

The heart dataset was subjected to one Hot encoding to convert categorical data into a numerical format. Thereafter, the dataset went through SMOTE to correct the bias in the dataset. Then features extraction was done using Autoencoder in order to increase the performance of the Model. The Autoencoder has input layer that consists of the same number of neurons as features (13) used in the dataset. Each neuron represents one features such as has the same number of neurons one features (age, sex, cp, trestbps, chol, fbs, restecg, thalachh, 'exang, oldpeak, slope, ca, thal. The input data is fed into the encoder, which processes it through its layers to produce the encoded representation in the latent space.The Encoder has three layers with three neurons as follows: 128, 64 and 32 for first layer, second and final layers respectively. The choice of neurons used depends on the complexity of the dataset and the level of compression. The decoder also have three neurons such as 64 ,128 and 13 for first, second and output respectively and uses sigmoid activation to reconstruct the original image as seen in Figure 4.

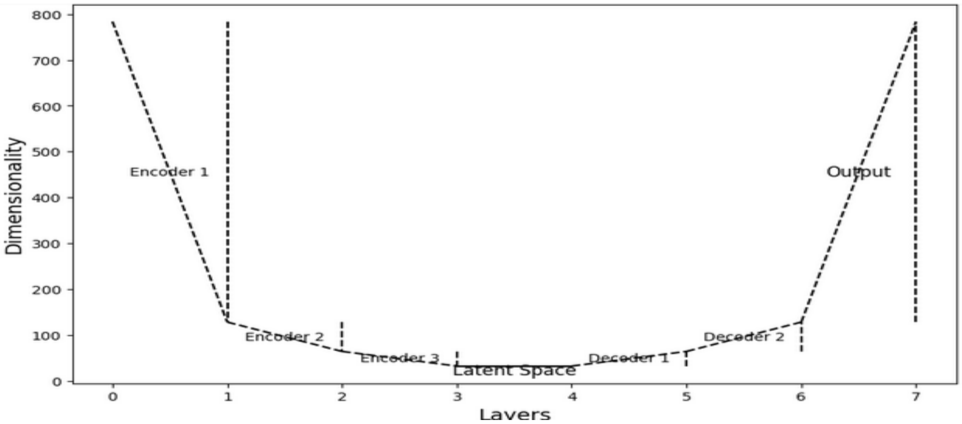


Figure 4: Autoencoder Model for feature extraction

3.2 Experiment Results and Discussions

The Autoencoder model developed was able to reconstruct the original image and the image reconstructed is similar to the original image as seen in Figure 3 which implies the autoencoder model learn effectively. The MSE and RMSE graph plot shows that as the epoch is increase the MSE and RMSE value is reducing towards zero, this implies that the model performance is improving as the epoch values is increases as seen in Figure 5 & 6 . The Loss plot also revealed that the error values reduces as the epoch increases which shows that model error is reducing as the model improves on its prediction as seen in Figure 7.

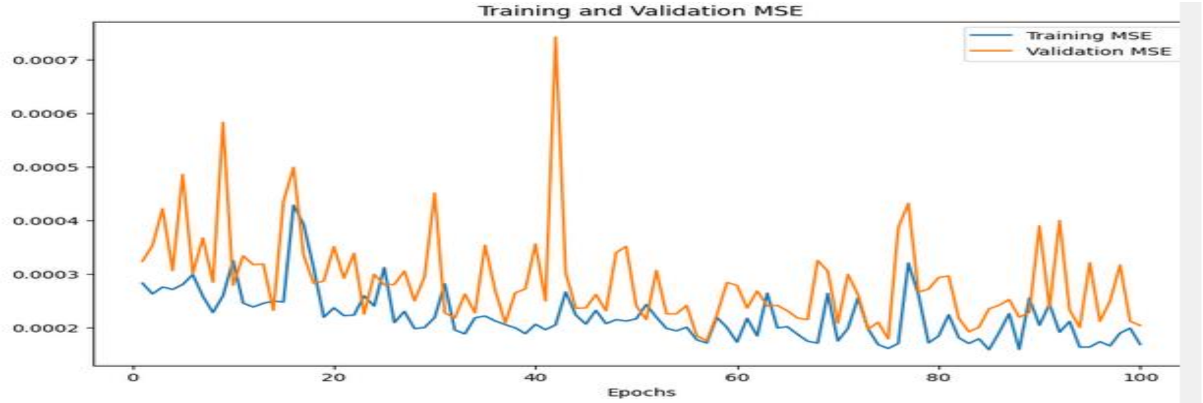


Figure 5: MSE for Autoencoder

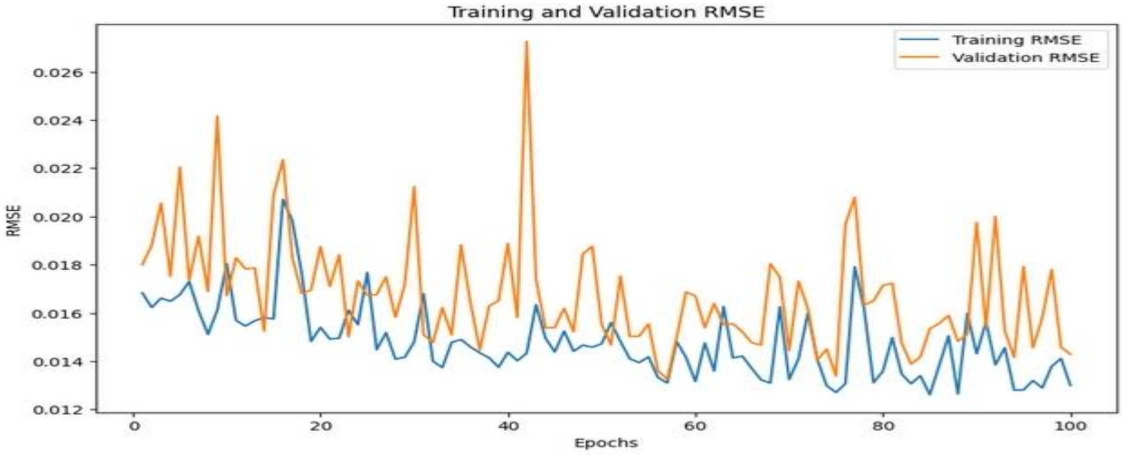
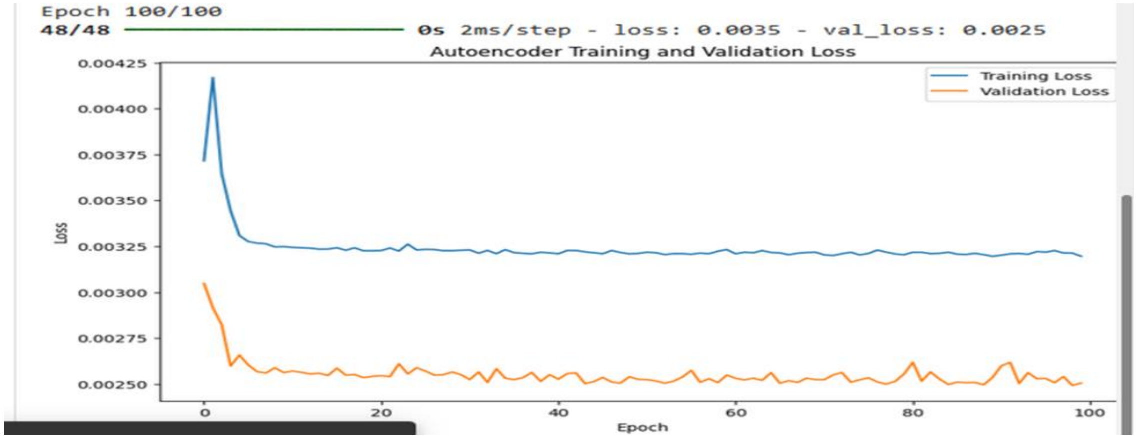


Figure 6 RMSE for Autoencoder



```
Epoch 100/100
48/48 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - loss: 0.0035 - val_loss: 0.0025
```

Figure 7: Loss value for Autoencoder

Table 2 revealed that the autoencoder achieved a very low MSE of 0.00031, indicating a good fit to the data and suggesting the HD predictive model accurately reconstruct the input features and perform well. The results also revealed a value of 0.0177 for RMSE indicating a good fit to the data and suggesting the HD predictive model accurately reconstruct the input features and perform well (Chai, & Draxler 2014). The results show that the model achieved a low MAE of 0.0120, indicating a good fit to the data and suggesting the HD predictive model accurately reconstruct the input features and perform well (Willmott & Matsuura, 2005). The autoencoder achieved a very high R2 of 0.9917, indicating a good fit to the data and suggesting the HD predictive model accurately reconstruct the input features and perform well (Kvalseth,1985). The results also revealed that model achieved high accuracy of 91%.

Table 2: Evaluation Results of Autoencoder used for feature extraction

Metrics	Scores
MSE	0.00031
RMSE	0.0177
R2	0.9917
MAE	0.0120
Accuracy	91

LightGBN Model

The result of the classification model done with LightGBM with confusion for the HD predictive model shows that LightGBN correctly predicted 180 people as not having HD (class 0), incorrectly predicted 8 as having HD (class1), 10 as not having HD (class 0) and correctly predicted 180 as having diabetes (class1) as seen in Figure 8

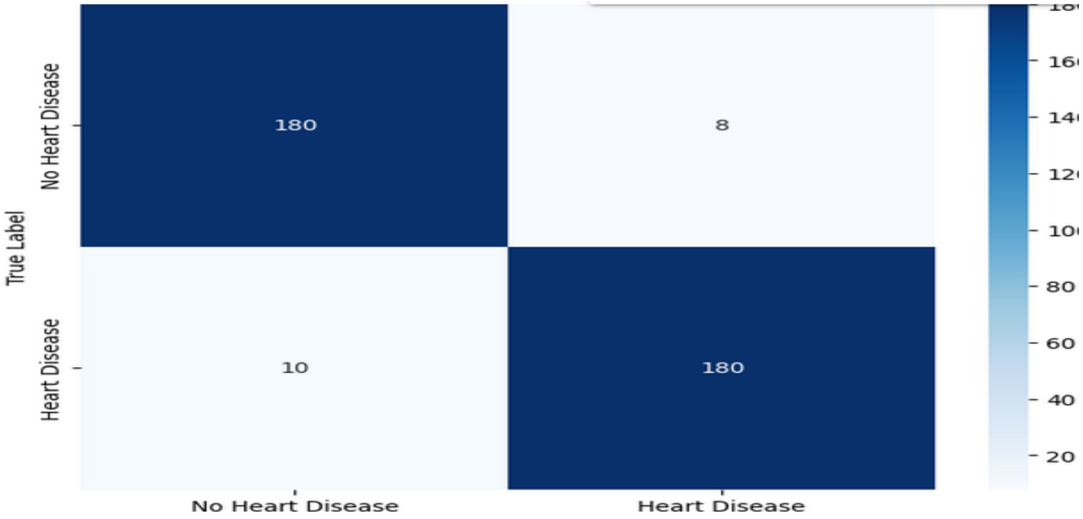


Figure 8: LightGBM confussion Matrix for the HD

The Results of the LightLGB ROC for HD prediction shows an AUC value of 0.99 which means that if a person with HD and one without HD is randomly picked from the dataset , Based on the result its clearly demonstrated that there is 99% chance that the test will give a higher score to the person who actually has HD as seen in Figure 9

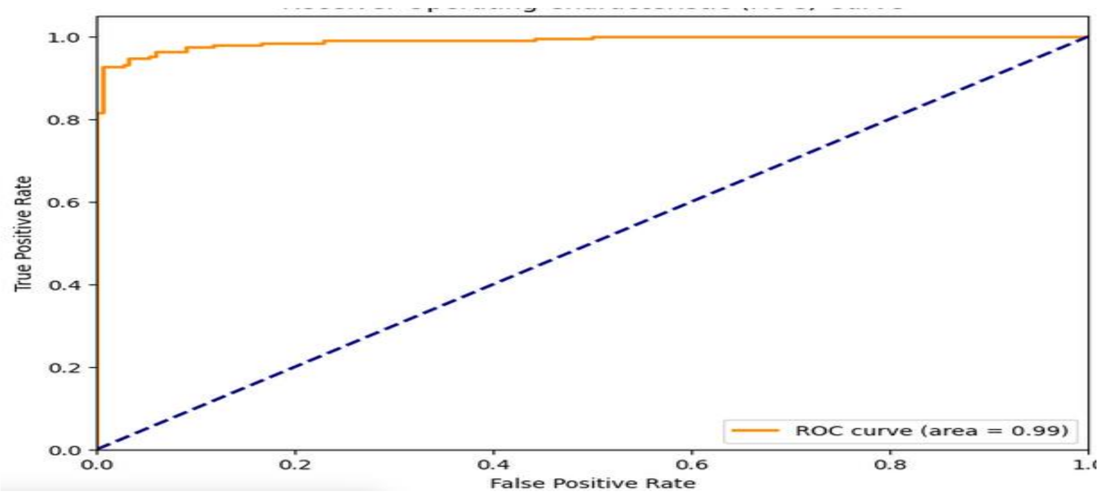


Figure 9: Light GBM ROC curve

Gradient Boosting Model

The result of the confusion matrix for the HD predictive model shows that Gradient Boosting correctly predicted 172 people as not having HD (class 0), incorrectly predicted 16 as having HD (class1), 19 as not having HD (class 0) and correctly predicted 171 as having diabetes (class1) as seen in Figure 10

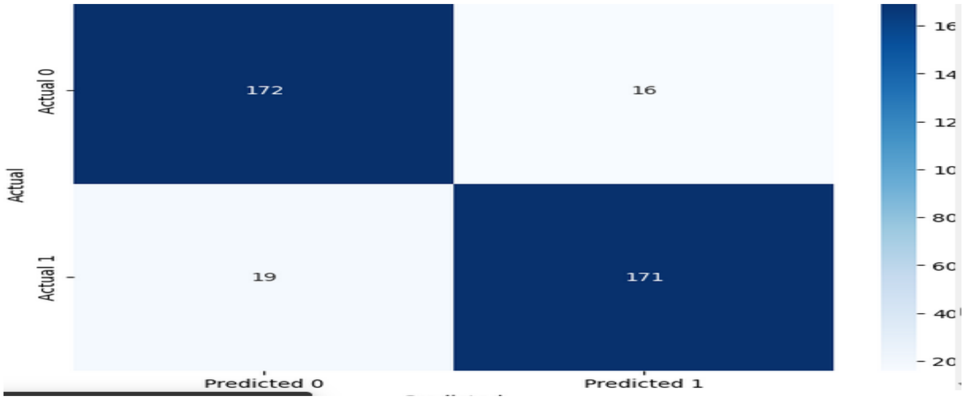


Figure 10 Gradient Boosting confusion Matrix

The Results of the Gradient Boosting ROC for HD prediction shows an AUC of 0.96 which means that if a person with HD and one without HD is randomly picked from the dataset, Based on the result its clearly demonstrated that there is 96% chance that the test will give a higher score to the person who actually has HD as seen in Figure 11

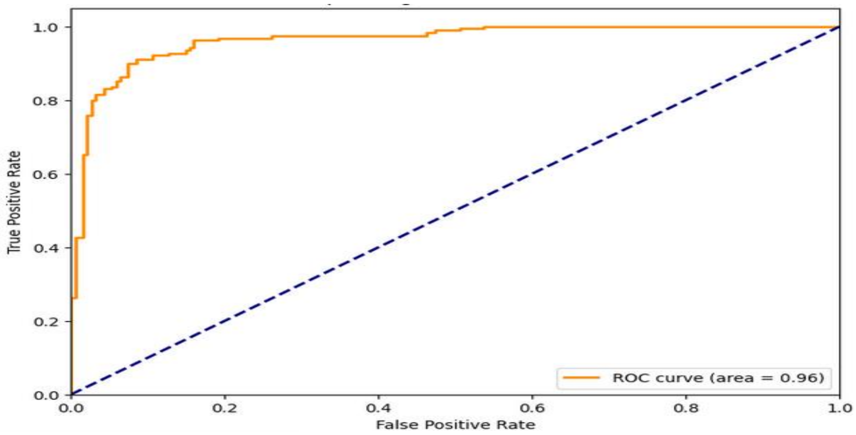


Figure 11 Gradient boosting ROC curve

The bar plot Result for the Heart Disease prediction

The result of bar plot show a clearer picture of the relative risk of HD based on chest pain. The bar plot showing percentages provides a clearer picture of the relative risk of heart disease based on chest pain type. The model revealed that there's typical and atypical angina that have higher percentage for patients with HD which shows the importance in heart prediction. The target variable =1 suggests a strong association as seen Figure 12

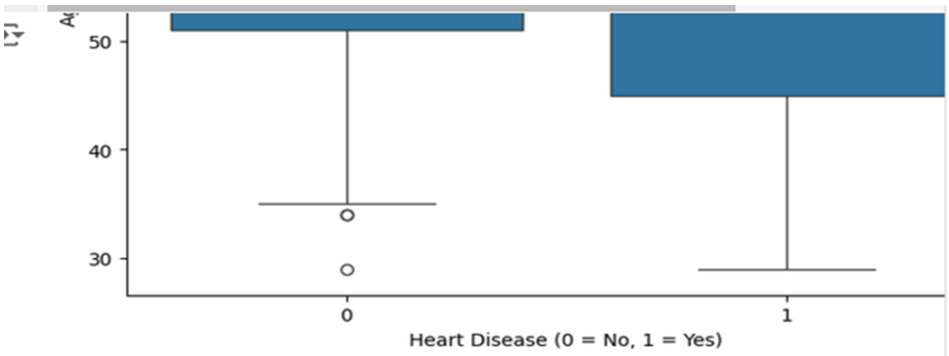


Figure 12 Box plot for the HD Prediction

The count plot for the heart disease

The result of count plot revealed that chest pain (CHP) has four categories namely: chp=0, chp =1, chp=2 and chp=3. Figure revealed that for the category cp=0,it indicates a strong relationship between typical angina and the presence of HD. For cp=1 shows an increased count for patient with HD . For cp=2 category a higher count for patients without HD and its represents chest pain not related to heart issues. The cp-3 categories have a mixed distribution, as some patients with HD might not experience chest pain as seen in Figure 13

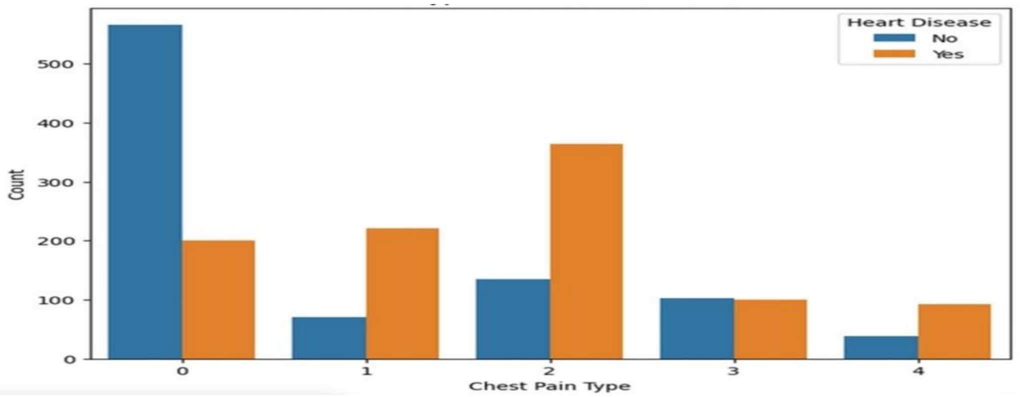


Figure 13: Count Plot for Chest pain versus hearth disease

The Features Importance

The results of features importance of the HD plotted revealed that Resting electrocardiographic results (restecg) and Resting blood pressure (Tresbps) have high importance score of 431 and 412 respectively, it suggests that these two factors are strong indicators of HD risks. The results of the Serum cholesterol (chol) has a low importance score of 18 it suggests that these factor is a weak indicator of HD risks as seen in Figure 14

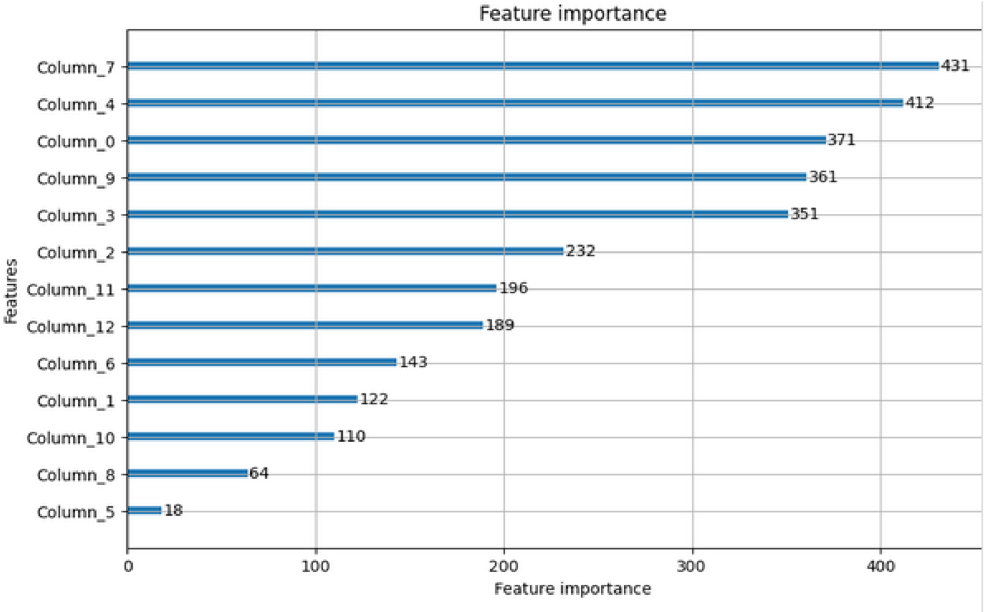


Figure 14: Heart Disease feature importance

Discussions

Table 3 revealed that LightGBM model has a high precision of 96% meaning its good in avoiding false positive that is incorrectly classifying someone as having HD when they do not. The result revealed that the model has a good recall result of 95% indicating it identifies most of the actual HD cases. The F1 score of 95% reflects a good balance between precision and recall. The model give a specificity result of 96% meaning that the model correctly identified 96% of individuals who did not have HD. The accuracy result of 95% suggests that the model performed well in identifying both individuals with and without HD. This shows that the model overall ability to make correct prediction. Table 4 revealed that gradient Boosting model has a high precision value of 91%. The recall result of 90% indicating it equally good in recognizing HD cases. The F1 score of 91% reflects a good balance between precision and recall. The model give a specificity result of 91% meaning that the model correctly identified 91% of individuals who did not have HD. The accuracy result of 90%.

Table 3: Evaluation Results of LightLGBM

Metrics	**Scores**
Specificity	95.7%
Precision	95.7%
Recall	94.7%
F1	95.2%
AUC	0.99
Accuracy	95

Table 4: Evaluation Results of Gradient Boosting

Metrics	**Scores**
Specificity	91.7%
Precision	91.7%
Recall	90.%
F1	91.%
AUC	0.96
Accuracy	90

Comparison of the two Models

The comparison results of the model shows that LightGBM outperform the boosting gradient as shown in Figure 15 and 16 in term of accuracy and ROC.

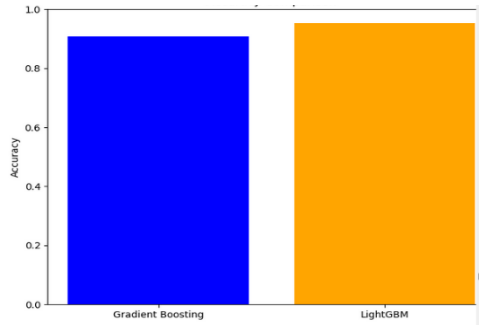


Figure 15: Accuracy compassion of models

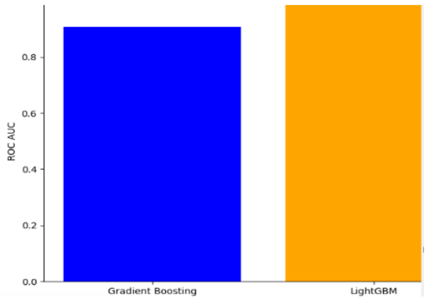


Figure 16: ROC compassion of models

The study of Heart Disease predictive model using autoencoder with LightGBM and Gradient boosting developed in this paper is compared with the results of the study carried out by Folorunsho et al (2024) on prediction of Diabetes risks using Autoencoder with Explanable AI the model of the study achieved accuracy of 89.5 % the study outperform the study with accuracy of 95% . Also compared with study carried out by Divya et al. (2024); Kardam et al. (2024) on prediction of HD where the model achieved an accuracy results of 93.5%. and 81% respectively. Our model outperform the model with an accuracy of 95%.

4.0. Conclusion

The goal of this study is to examine the strength of ML in prediction using Autoencoder with LightGBM and Gradient Boosting to predict the HD with 13 predicting factors based on the dataset collected from kaggle.com repository. The study developed autoencoder with LightGBM and Gradient Boosting for HD predictive model provides a promising approach for HD prediction.One Hot encoding and SMOTE were used for preprocessing of the dataset The validating metrics used for the Autoencoder model revealed the model perform well with high values for accuracy and R2 and low values for MSE, RMSE and MAE metrics. The validating metrics used for the LightGBM and Gradient Boosting models revealed that the two models perform excellently with high values for Specificity , Precision, Recall, F1, AUC and Accuracy. The study concluded that LightGBM perform better than Gradient Boosting based on the metrics used for validation. Also the study revealed that there is relative risk of HD based on chest pain type as clearly seen in the bar plot and count plot developed for HD prediction. So also that Resting electrocardiographic results (restecg) and Resting blood pressure (Tresbps) are strong factors to HD prediction. The model is recommended to the health sector management to guide their decision-making. Its potential integration with predictive model and clinical validation will assist greatly to improve the HD diagnosis and prevention .

5.0 Recommendation

- Extensive works could be done with more validating metrics and more deep learning techniques could be employed to increase the accuracy on the model.
- Different clinical dataset for HD could be used.
- Classification threshold can be varies in order to improve the recall

Acknowledgements

We sincerely appreciate all authors whose works is cited in this paper

References

Akare, U., Gani, U. A., Bhongade, A., Mure, D., Chatterjee, M. & Ramteke, V. 2024. Heart disease prediction system using machine learning. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),* 13(3), 92-97. DOI: 10.17148/IJARCCE.2024.13315

Abass, O.A., Alaba, O.B. & Ebo, I.O. 2020 Artificial Neural Network Approach to Determine Cerebrovascular Accident Risk Factors among Patients in Ogun State, Nigeria Nigerian Journal of Scientific Research, 19 (3); abu.edu.ng; ISSN-0794-0319

Bharani, B. R., Manjunatha, S, Vijayalakshmi, R. Y. & Preethi, S. 2024. Heart disease prediction using effective machine learning techniques. *International Journal for Multidisciplinary Research (IJFMR),* 6(2), 1-8.

Bombale, G., Pawar, A., Bhole, R., Pawar, A. & Duble, Y. 2024. Heart disease prediction using machine learning. *International Research Journal of Modernization in Engineering Technology and Science (IRJMES),* 6(10), 2908-2913.

Divya, N., Riyazuddin, M., Ahad, A., Vulapula, S. R., Manjula, A. & Sirajuddin, M. 2024. Predicting heart disease using machine learning and IoT techniques. *Oncology and Radiotherapy* 18(9), 1-11.

Gour, S., Panwar, P., Dwivedi, D. & Mali, C. 2022. A machine learning approach for heart attack prediction. *Intelligent Sustainable Systems Springer,* 741–747. https://doi.org/10.1007/978-981-16-6309-3_70

Kardam, H., Srishti & Deepanshu 2024. Heart disease prediction using machine learning. *Journal of Novel Research and Innovative Development (JNRID),* 2(6), 1-12.

Logabiraman, G., Ganesh, D., Kumar, M. S., Kumar, A. V. & Bhardwaj, N. 2024. Heart disease prediction using machine-learning algorithms. MATEC Web of Conferences, pp. 1-9 https://doi.org/10.1051/matecconf/202439201122

Pradip, M. P. & Atharva, J. 2024. Heart disease prediction using multiple machine learning algorithms. *Advances in Robotic Technology*, 2(1), 1-5.

Saha, S., Rahman, M., Suki, T. T., Alam, M., Alam, S. & Haque, M. A. S. 2024. Heart disease prediction using machine learning algorithms: Performance analysis. *3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pp. 1-7.

Shreya, K. C. & Sowmya, K. S. 2024. Heart disease prediction. *International Journal of Research Publication and Reviews*, 5(1), 5192-5198.

World Health Organization. 2021. Cardiovascular diseases (CVDs). Retrieved from [WHO website link].

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Johnson, J. M., & Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.

Chai, T., & Draxler, R. R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.

Folorunsho O., Amoo O. G., Odufuwa1 T. T., Ochidi I., Mogaji S. A. & Faboya, O.O. 2024. Prediction of Diabetes Risk Using Autoencoder with Explainable Artificial FUOYE Journal of Pure and Applied Sciences Vol 9(3) ISSN: 2616- 1419.

Willmott, C. J., & Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Kvalseth, T. O. 1985. Cautionary note about R 2. The American Statistician, 39(4a), 279-285.

Meng K. G., Finley, Q., Wang, T., Chen,T., W., Ma, W., Ye, Q., & Liu, T. Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Ogunsanwo G .O (2024) Predictive Model for Health Insurance Cost using Self-Organizing Maps and XGBOOST FUDMA. Journal of Sciences (FJS), 8(6), 354-362.