Open Journal of Physical Science (OJPS)
ISSN: 2734-2123
Article Details:
DOI: 10.52417/ojps.v4i1.457
Article Ref. No.: OJPS0401002-457
Volume: 4; Issue: 1, Pages: 19- 33 (2023)
Accepted Date: 3rd February, 2023
© 2023 Mohammed & George

Open Journals Nigeria (OJN)
Open Access | Bi-annual | Peer-reviewed
www.openjournalsnigeria.org.ng
editorial@openjournalsnigeria.org.ng

RESEARCH ARTICLE **OJPS0401002-457**

IDENTIFICATION AND MITIGATION OF BIAS USING EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR BRAIN STROKE PREDICTION

Mohammed, K. & *George, G.

Department of Computer Science, Faculty of Computing and Applied Sciences, Baze University, Jabi Abuja, Nigeria

***Corresponding Author Email:** gilbert.george@bazeuniversity.edu.ng*

ABSTRACT

Stroke is a time-sensitive illness that without rapid care and diagnosis can result in detrimental effects on the person. Caretakers need to enhance patient management by procedurally mining and storing the patient's medical records because of the increasing synergy between technology and medical diagnosis. Therefore, it is essential to explore how these risk variables interconnect with each other in patient health records and understand how they each individually affect stroke prediction. Using explainable Artificial Intelligence (XAI) techniques, we were able to show the imbalance dataset and improve our model’s accuracy, we showed how oversampling improves our model’s performance and used explainable AI techniques to further investigate the decision and oversample a feature to have even better performance. We showed and suggested explainable AI as a technique to improve model performance and serve as a level of trustworthiness for practitioners, we used four evaluation metrics, recall, precision, accuracy, and f1 score. The f1 score with the original data was 0% due to imbalanced data, the non-stroke data was significantly higher than the stroke data, the 2nd model has an f1 score of 81.78% and we used explainable AI techniques, Local Interpretable Model-agnostic Explanations (LIME) and SHapely Additive exPlanation (SHAP) to further analyse how the model came to a decision, this led us to investigate and oversample a specific feature to have a new f1 score of 83.34%. We suggest the use of explainable AI as a technique to further investigate a model’s method for decision-making.

Keywords: *Brain Stroke Prediction, Explainable AI, Machine Learning, LIME, SHAP*

INTRODUCTION

With the aid of Artificial Intelligence (AI), we have seen unbelievable progressions in the medical profession (Pamungkas *et al.*, 2022) (Chen *et al.*, 2022; Kwon *et al.*, 2019a, 2019b; M *et al.*, 2020)from cancer predictions using machine learning to covid-19 detection using deep learning. We may use data mining techniques to discover trends in the dataset now that there is a collection of medical records that have been labelled. Thanks to such assessments, medical experts can now forecast the prognosis of any medical problem with accuracy. Better healthcare conditions and decreased medical costs are the results of this. The use of data mining techniques in medical records has had a considerable impact on the fields of healthcare and biomedicine.(Chen *et al.*, 2022; Kwon *et al.*, 2019a, 2019b; Lu *et al.*, 2020; M *et al.*, 2020; Maharjan, 2020). This enables doctors to detect the beginning of a disease at an earlier stage. Finding the main causes of stroke and its risk factors is something we are very interested in.

Numerous studies have looked at the significance of people's medical histories and lifestyle choices on the likelihood that they would have a stroke (Islam *et al.*, 2022; Khosla *et al.*, 2010; Pamungkas *et al.*, 2022; View of Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients, n.d.). The likelihood of stroke is currently predicted using machine learning models as well. As well as implementing explainable AI, our goal in this study is to employ machine learning algorithms to forecast a person's risk of stroke based on their lifestyle.

According to the World Health Organization, cardiovascular diseases such as ischemic heart disease and stroke are the leading cause of death worldwide (View of Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients, n.d.).It is not new to research, explainable AI. Since the early 1980s, when CS Pierce's deductive reasoning was used in those systems, there have been reasoning structures to support expert systems (Gunning *et al.*, 2019).

AIM AND OBJECTIVES

When blood flow to the brain is impaired, a dangerous and even fatal medical emergency known as a stroke can happen. It has a serious physical, psychological, and financial impact on people, families, and society as a whole and is a leading cause of disability and mortality globally. Our research aims to highlight important features that will aid and improve the ability of the machine learning model to efficiently predict stroke. We perform a prediction task and compare two Explainable AI techniques for predicting brain stroke and identifying important.

The objectives of this study are:

- To use machine learning for the prediction of brain stroke
- To use Explainable AI techniques on the brain stoke dataset
- To investigate bias using Explainable AI

REVIEW OF RELATED LITERATURE

In the study by Khosla *et al.* (2010), four machine-learning techniques were used for the detection of stroke, naive base, J48, k-nearest neighbour, and random forest. The accuracy of naive Bayes was 95.7, while 99.8% was the accuracy of J48, K-nearest neighbour, and random forest.

There has been researched in teaching AI to explain itself. Salient maps were used to show the location of where a network was used to decide its outcome, however, this gives little information about why an image was classified or misclassified(Kim *et al.*, 2018). Autoencoders' neural networks can use learned representations to apply prior knowledge about structures. Clough et.al developed a variational auto-encoder (VAE) to provide naturally interpretable concepts (Kim *et al.*, 2018). This method could be extended and combined with other networks for the project

Several papers have used Explainable AI in brain stroke prediction. The authors (Islam *et al.*, 2022) used an explainable AI LIME technique to predict stroke using EEG signal but research was limited due to lesion location and few cortical electrodes. The authors in (View of Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients, 2022.) used various structured and unstructured data and demonstrated transparency to a certain level and feature importance.

The authors (Pamungkas *et al.*, 2022) performed explainable AI using common features and discovered women had a higher risk of stroke compared to men, and those with late detection of heart diseases and hypertension still have a high risk of stroke.

MATERIALS AND METHODS

RESEARCH DESIGN

We proposed the following steps as our methodology for investigating bias.

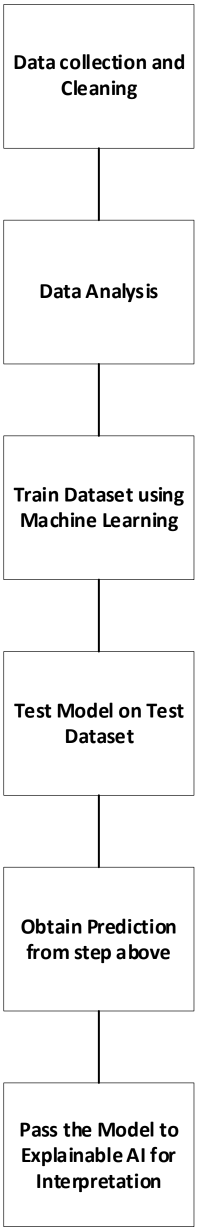

Data collection and Cleaning
Data Analysis
Train Dataset using Machine Learning
Test Model on Test Dataset
Obtain Prediction from step above
Pass the Model to Explainable AI for Interpretation

Figure 1: The research steps of the proposed method

The above figure displays the steps used in the proposed method, we start by cleaning the dataset, then we performed data analysis and converted the data into a numerical format for easy prediction, we then passed the dataset to the machine learning algorithm namely (Logistic regression, random forest, and Naive Bayes). Afterward, we created the model and performed testing on the unseen dataset, took the accuracy readings, and then post hoc explainability was used using SHAP and LIME techniques.

Data collection

Our research data was obtained from Kaggle from this website https://www.kaggle.com/code/ruthvikpvs/stroke-data-analysis-and-prediction/data and contained the following attributes and values:

- Age: The age of the patient's
- Gender: Could be 'Female', 'Male', or other
- Hypertension: the value is 1 if the patient has hypertension and 0 if the patient does not have hypertension
- Heart disease: a value of 1 if the patient has a heart disease or 0 if the patient does not have any heart disease
- Ever-married: A value of 'Yes' if the patient has been married or 'No' otherwise
- Worktype: Could be either 'self-employed', 'GovtJob', 'private', 'never worked' or 'children'.
- Type of Residence: Either "Urban" or "Rural".
- Smoking status: unknown', 'never smoked', 'formerly smoked', 'or' 'smokes'.
- BMI: body mass index
- Average glucose level: Average glucose level in the blood
- Stroke: '1' if the patient has a stroke or '0' otherwise.

Data analysis

The data contained 4981 entries; each field contained non-null values (all contained data). 248 entries were positive for store outcomes (about 4.98%) and 4733 entries were negative for stroke outcomes (about 95.02%)

Data formatting

For easy prediction and analysis of data, we changed the fields with "object" data type to number label values. The fields changed were: Gender, ever married, Work type, residence type, and smoking status. The new field values were:

- Gender: 0 for male and 1 for female
- Ever_married:0 for yes and 1 for No
- Work_type: 0 for private, 1 for self-employed, 2 for govt job and 3 for children
- Residence type: 0 for urban and 1 for rural
- Smoking_status: 0 for formerly smoked, 1 for never smoked, 2 for smokes and 3 for unknown

Data splitting

In our dataset, we stored the features needed to predict stroke as features and the actual value we want to predict as labels.

We stored the dataset excluding the stroke (outcome) column in our features variable 'x'. We stored the stroke outcomes in our label variable 'y'.

We performed an 80-20 test split on the data. We stored 80% of the data in a train set then trained the machine learning model and stored 20% for testing to test the model on unseen data.

Machine learning models used on the stroke dataset

Logistic regression

Logistic regression allows dichotomy or binary outcomes with 2 mutually exclusive levels to be analysed. It allows the use of categorical, continuous, and multiple predictors (Zhang & Han, 2020)

Naive Bayes

Naive Bayes uses the Bayes rule with the strong assumption that given the class, attributes are conditionally independent. We implemented the Gaussian naive Bayes (Rish, n.d.)

Random Forest

Random forests are combined tree predictors. An individual tree is determined by the values of random vectors sampled independently with the same distribution for every tree in the forest (Rigatti, 2017)

Evaluation metrics.

To assess the machine learning models, some common metrics are used: Recall, Precision, Accuracy, and F1 score. To get these metrics, after testing the models, the true positives, false positives, true negatives, and false negatives are used. True positives are exampling the system classified as positive and it was positive, a true negative is exampling the system classified as negative and it was negative, false positives are examples of the system classified as positive when it was negative, the false negatives are examples the system classified as negative when it was positive for all positive examples. Precision considers all positive samples (stroke) that are classified as positive either correctly or incorrectly. Recall shows the percentage of examples with a stroke (positive) that were correctly identified as positive Precision and recall help evaluate a model when dealing with imbalanced data. The F1 score is the harmonic mean of recall and precision. It uses the combination of precision and recall and weights them for overall performance.

$$recall = \frac{TP}{TP+FN'} \quad (1)$$

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

$$F - Measure = 2\,\frac{precision * recall}{precision + recall'} \quad (3)$$

$$Accuracy \ = \frac{TN+TP}{TN+TP+FN+FP} \qquad (4)$$

TP= true positive,TN=true negative FN=false negative,FP= false positive

Explainable AI techniques

After training and testing our machine learning model using random initialization. Explainable AI can be classified into two types, post-hoc explainability, and inherent explainability. The post-hoc explainability happens after the model has been trained or a prediction has been made. They are usually used with complex models. Examples are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapely Additive Explanation). Inherent explainability can be explained simply by looking at the data (out of the box), without the use of models or libraries, for example, a linear regression model where the price of a house goes up as the size of the house increases.

In this study, we implemented the SHAP and LIME techniques.

Table 1: Proposed Methods to be used in the investigation of Bias using explainable AI

Method	Dataset	XAI Technique
Model1 (M1)	Imbalance data	SHAP and LIME
Model2 (M2)	Balanced data	LIME
Model3 (M3)	Better performance using explainable AI	LIME

M1 - We will use M1 as the first model on the original dataset which is imbalanced. We will use SHAP and LIME techniques to explain the data so as for bias to be observed.

M2 - We will use M1 as the second model after oversampling the original data and will use the LIME technique on the model. After using the explainable AI technique to observe the features. We will further investigate and mitigate bias to obtain a better model.

M3 - We will use M3 as the third model to show how our investigation using explainable AI, improved the model's performance.

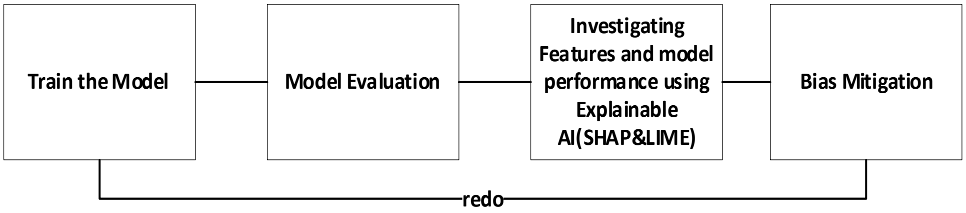

Train the Model
Model Evaluation
Investigating Features and model performance using Explainable AI(SHAP&LIME)
Bias Mitigation
redo

Figure 2: The process of investigation for bias using explainable AI

Algorithm

Step 1: Training the model

Step 2: Evaluating the model

Step3: Investigations for bias using Explainable AI

Step 4: If model is biassed go to step 1-3

Step 5: Else End

RESULTS

The ML models were implemented using TensorFlow and Python. To analyze the data, we observed how the features correlate with the output (stroke). We obtained the result in the figure after running the correlation between the input data and output data.

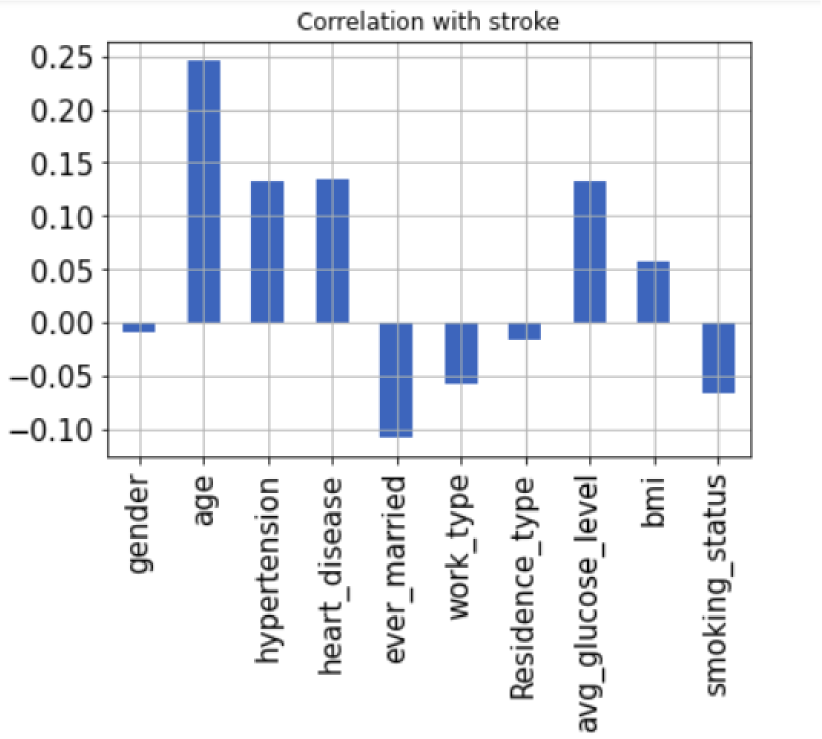


Figure 3. The feature correlation with our model

Model 1 (M1) Evaluation of the Imbalance Dataset

After training and testing the models, the logistic regression accuracy was 95.1, the random forest model had an accuracy of 94.78 and the Gaussian naive Bayes had an accuracy of 87.36. Therefore, the logistic regression model was used in the Explainable AI analysis. Even though the logistic regression model had an accuracy of 95%, it performed poorly because it predicted all the test data as negative and none as positive. This is because of the imbalance in the dataset. It could not predict any true positive therefore the precision, recall, and accuracy had a score of 0. Both SHAP and LIME techniques were used to investigate the imbalance dataset.

SHAP techniques on the imbalanced dataset

Feature importance

The plot below shows the feature importance of the data set calculated by SHAP values. Shapley values are utilised in cooperative game theory (two or more factors used in a strategy to reach a desired result) which involves a fair distribution of gains and costs to various actors working together. It is a simple understanding of the model. Important features are those with huge shapely values.

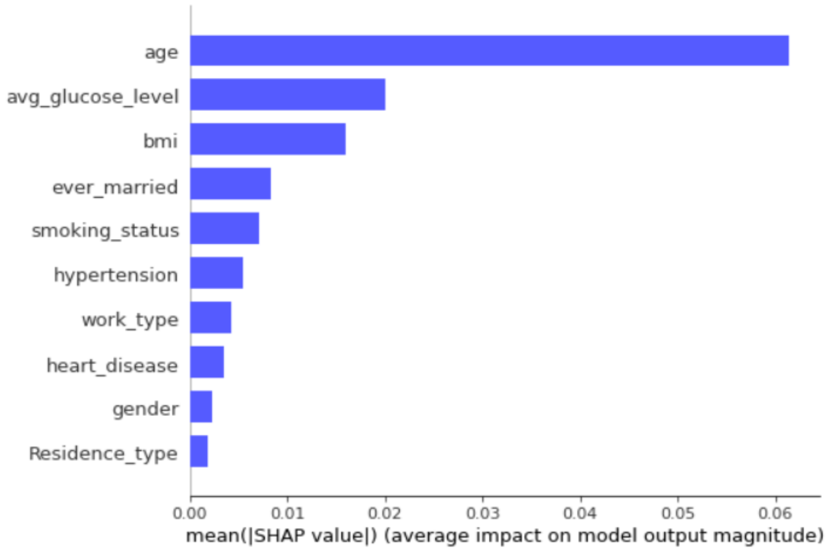


Figure 4: Feature importance based on SHAP values

Summary plot

The summary plot below integrates feature importance with feature effects. Each point is a shapley. The y-axis shows the feature anthem-axis showing the shapley value of each instance.

The average glucose level has a high shapley value range.

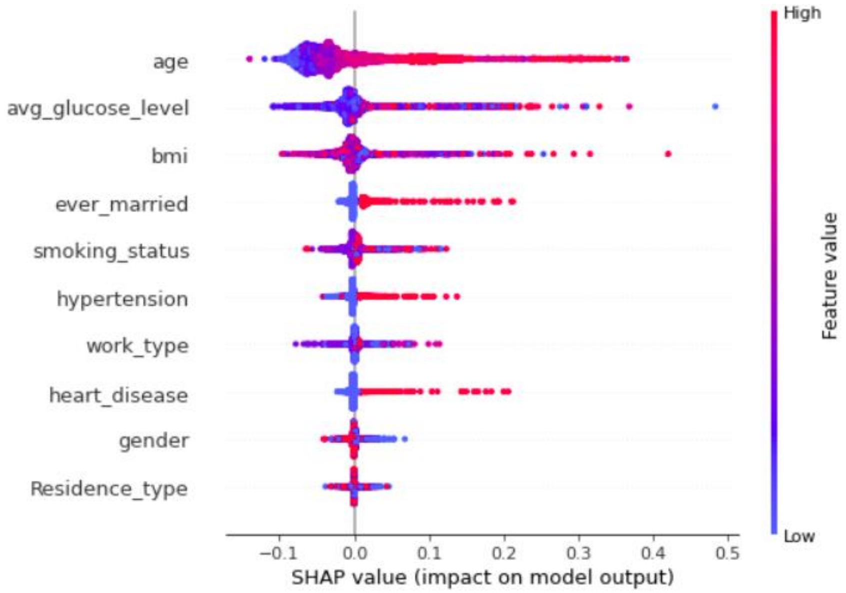


Figure 5: Feature importance with feature effects

The dependence plots

The dependence plot shows a scatter plot which indicates the impact of a single feature on the predictions the model has made. In the figure 6 below, the average glucose level increases significantly when the age reaches 40. Each dot is a prediction row in the dataset, the x-axi represents the features value and the y-axis represents its shapely value.

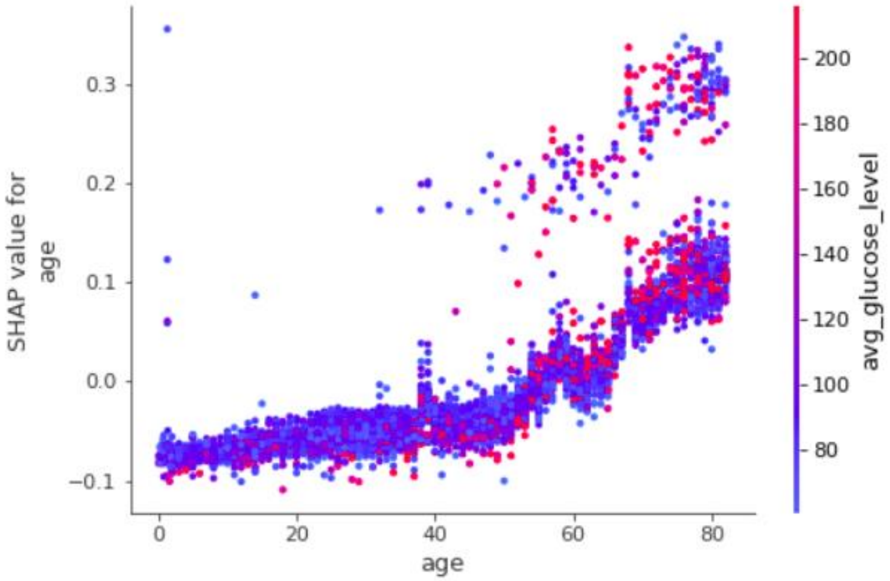
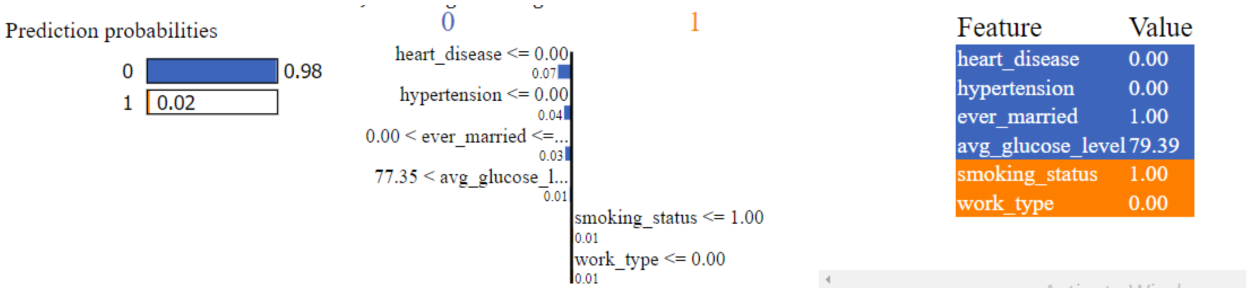


Figure 6: The dependence plot that indicates the effect of a single feature on the predictions the model

LIME Technique for Explainable AI

Positive prediction

The figure below shows a correct prediction of no stroke and shows the features that made the system come to that conclusion

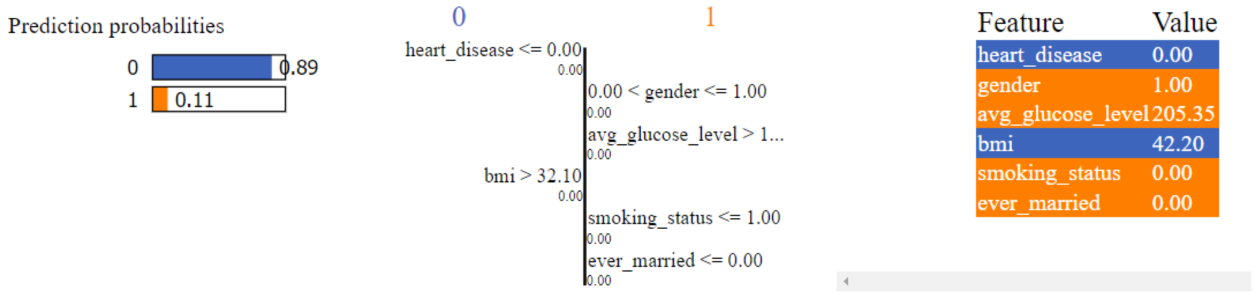


Feature	Value
heart_disease	0.00
hypertension	0.00
ever_married	1.00
avg_glucose_level	79.39
smoking_status	1.00
work_type	0.00

Figure 7: Positive prediction of LIME for M1

Negative prediction

The figure below shows how an output that should be '1' (stroke) was explained as non-stroke. This explanation shows the problem in the dataset or why it has that output which is one of the benefits of explainable AI.



Feature	Value
heart_disease	0.00
gender	1.00
avg_glucose_level	205.35
bmi	42.20
smoking_status	0.00
ever_married	0.00

Figure 8: Negative prediction of LIME for M1

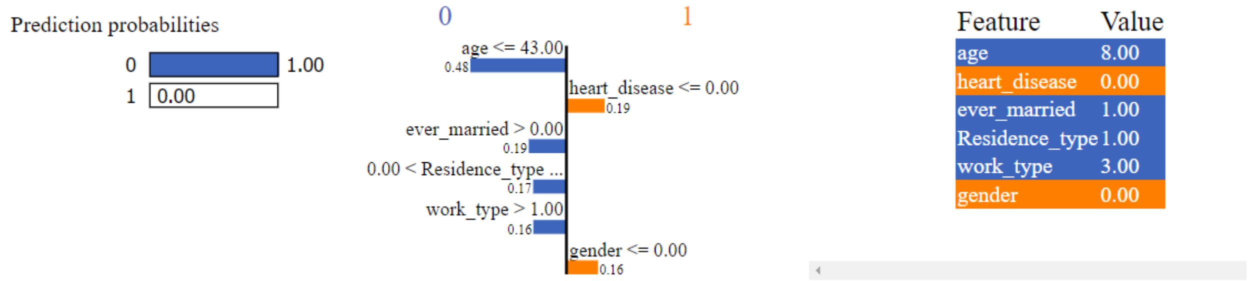
This can be used to observe that there exists bias in the dataset. The figure shows that one of the probabilities the system used to come to its conclusion was 'gender;. One reason for this is due to bias/unbalanced dataset. A professional may help spot this problem based on the visualisation to discard that particular production.

Model2 (M2): Balancing the dataset and Model Evaluation

Accuracy could not be used as a measure of accuracy due to the unbalanced dataset. To solve the issue of the imbalance dataset, we balanced the data by oversampling the positive ('1' for stroke) outcomes. This increased the dataset from 4981 to 9466. The positive values increased from 248 to 4733 This gave us a precision score of 80.2 and recall score of 83.42 higher f1 score of 81.78%. Now the model predicts positive outcomes. We performed an explainableAI technique (LIME) on the new data for analysis.

Positive outcome

In the figure below, the probability of having a stroke is 0% for that patient's data. Based on this figure we can say the young age had an impact in making it '0', the patient has had a heart disease but it shows the other features had importance in this outcome.

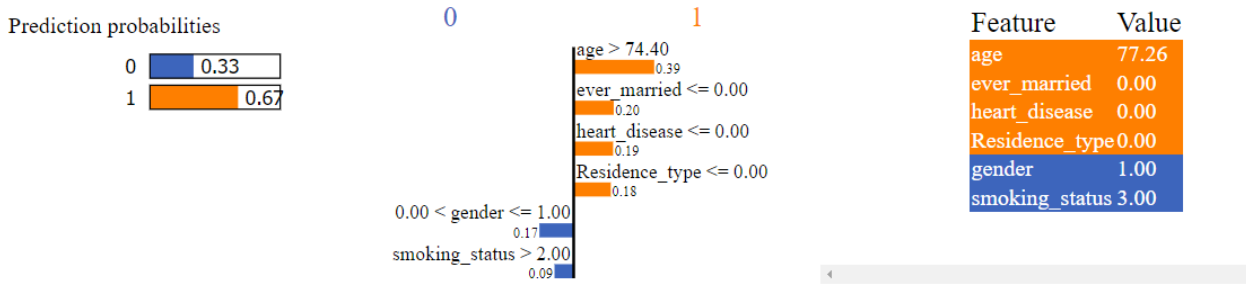


Feature	Value
age	8.00
heart_disease	0.00
ever_married	1.00
Residence_type	1.00
work_type	3.00
gender	0.00

Figure 9: Positive prediction of LIME for M2

Negative outcome

The figure below explains how the system came to its prediction of stroke with a 67% probability, an age greater than 74 (the patient was 77). It explains that the 'ever_married', 'residence type', and 'heart disease' was also significant in making this decision.



Feature	Value
age	77.26
ever_married	0.00
heart_disease	0.00
Residence_type	0.00
gender	1.00
smoking_status	3.00

Figure 10: Negative prediction of LIME for M2

Observation after oversampling

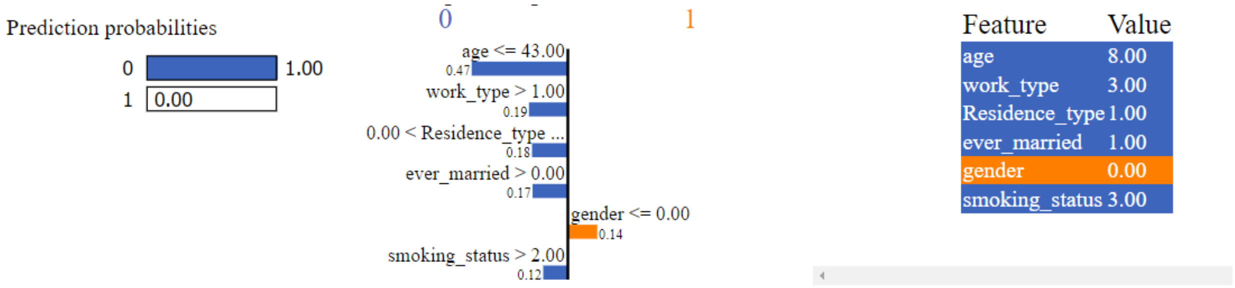
After observing the above data, we noticed the feature of heart disease had an impact in making the outcome positive even though the example had a negative (no) for heart disease.

Model3 (M3): Investigating Feature and Model Evaluation

Based on the previous task, we noticed the model used heart disease to lean towards the positive (stroke) decision when the patient has not had any heart disease. To find out why we looked through the dataset. We found that for heart disease, there were only 455 examples of negative and 9011 examples of positive. So we oversampled based on the heart disease to have equal (positive and negative) data for heart disease. After this, we had a new dataset size of 17994. Our new precision score was 80.51 and recall was 86.38 and the f1 score was 83.34

Negative outcome (no stroke) investigation

Due to oversampling the heart_diseases feature, we can see from figure [] below that where a patient did not have heart_disease is no longer a measure in leaning towards a positive output for stroke.

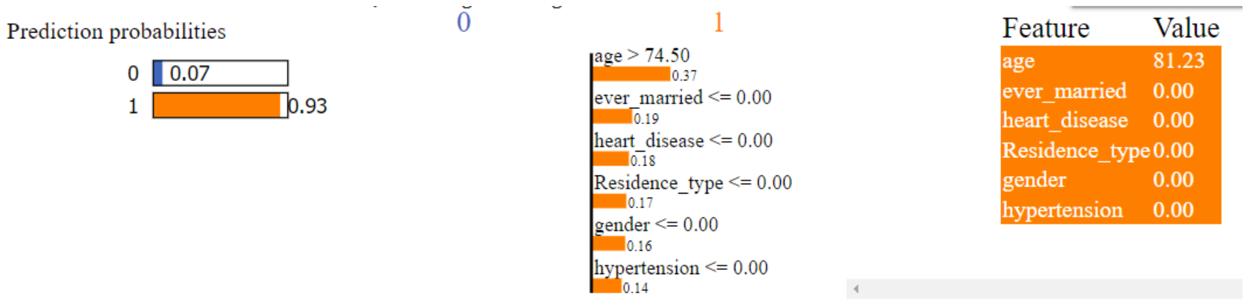


Feature	Value
age	8.00
work_type	3.00
Residence_type	1.00
ever_married	1.00
gender	0.00
smoking_status	3.00

Figure 11: Negative outcome (no stroke) investigation of LIME for M3

Positive outcome (stroke) investigation

In the figure below, we can see the improvement in the prediction,



Feature	Value
age	81.23
ever_married	0.00
heart_disease	0.00
Residence_type	0.00
gender	0.00
hypertension	0.00

Figure 12: Positive outcome (stroke) of LIME for M1

Table 2: Performance of the methods used for investigation of Bias

Method	Performance (f1 score)
M1	0%
M2	81.78%
M3	83.34%

M1 is the first model, this was used on the original dataset which was imbalanced. We used SHAP and LIME techniques to explain the data so the bias could be observed. The model classified all the data samples as non-stroke hence the reason for the f1 as 0%.

M2 is our second model, we oversample the original data and used logistic regression, naive Bayes, and random forest and used the highest score (logistic regression) to perform the LIME technique on the model. After using the explainable AI technique, we found the model was using some features in the wrong way. We further investigated this and tried M3.

M3 is the third model that was used to show how our investigation using explainable AI, improved the model's performance. We were able to improve our score from 81.78% to 83.34% just by investigating the features (age, gender, and heart disease) with explainable AI.

The copy of our source code could be accessed via the GitHub repository: https://github.com/khadijah13/ExplainableAI_stroke

DISCUSSION

Since there is no standard form to measure explainable AI, we suggest a certain step or approach where a medical practitioner could use the explainable AI to spot bias. This would help the practitioners know whether to trust the system to follow the prognosis. Here, we the AI developers can see that there is an imbalance in the dataset and can mitigate this issue. We Over-sampled the dataset which gave us a new size of 9466 examples. We tried explainable AI after oversampling and had better outputs. For example, we could see the system was leaning towards a score of 0 because the patient was 8 years old (less than 43). We also noticed that the system was using 'no heart disease' and ‘gender’ as a factor in predicting the patient has had a stroke. For this, we investigated the heart disease data and found that for heart disease, there were only 455 examples of negative and 9011 examples of positive data. We over-sampled based on heart disease to make this have equal data on heart disease. After this, we had a new dataset size of 17994. Our new precision score was 80.51 and recall was 86.38 and f1 score was 83.34. This improved the dataset. Our data still showed that one gender has a higher risk of stroke but on further investigation of the gender feature, we discovered that the gender features were also imbalanced. In the future, we suggest a way to balance that feature in a way that the other features are not affected or cause higher bias.

CONCLUSION

Based on our analysis, we discovered that due to the unbalanced dataset, the system learned to predict non-stroke but still had difficulty predicting stroke. We were able to show how explainable AI was used to spot bias in the dataset. We showed how Explainable AI could be of benefit to AI developers and practitioners by letting the AI developers see the impacts of certain features on the model. This can let the developers investigate a feature and mitigate the issue as shown in the example above. The model was biased towards one outcome over another due to an unbalanced dataset. We used oversampling to help mitigate this error. Using techniques like explainable AI shows why the data is making a decision. For example, the LIME technique showed us how it leans towards an outcome based on a feature’s value being lower or higher than a certain value. It also helps practitioners know whether to follow the AI’s decision for example, the practitioner would also see the system leads toward the score when the patient does not have any heart disease and can use that to understand how the AI decision is being made.

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

Chen, F., Sun, C., Yue, Z., Zhang, Y., Xu, W., Shabbir, S., Zou, L., Lu, W., Wang, W., Xie, Z., Zhou, L., Lu, Y., & Yu, J. (2022). Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **265**: 120355. https://doi.org/10.1016/J.SAA.2021.120355

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, **4**(37). https://doi.org/10.1126/SCIROBOTICS.AAY7120

Islam, M. S., Hussain, I., Rahman, M. M., Park, S. J., & Hossain, M. A. (2022). Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. *Sensors (Basel, Switzerland)*, **22**(24):. https://doi.org/10.3390/s22249859

Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 183–191. https://doi.org/10.1145/1835804.1835830

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & sayres, R. (2018). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)* (pp. 2668–2677). PMLR. https://proceedings.mlr.press/v80/kim18d.html

Kwon, H., Park, J., & Lee, Y. (2019a). Stacking Ensemble Technique for Classifying Breast Cancer. *Healthcare Informatics Research*, **25**(4): 283. https://doi.org/10.4258/HIR.2019.25.4.283

Kwon, H., Park, J., & Lee, Y. (2019b). Stacking Ensemble Technique for Classifying Breast Cancer. *Healthcare Informatics Research*, **25**(4): 283–288. https://doi.org/10.4258/HIR.2019.25.4.283

Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Znati, T., Mi, Q., & Jiang, J. (2020). Using machine learning to predict ovarian cancer. *International Journal of Medical Informatics,* **141**: 104195. https://doi.org/10.1016/J.IJMEDINF.2020.104195

Maharjan, A. (2020). *Machine Learning Approach for Predicting Cancer Using Gene Expression* (Doctoral dissertation, University of Nevada, Las Vegas).https://doi.org/10.34917/19412120

Pamungkas, Y., Wibawa, A. D., & Cahya, M. D. (2022). Electronic Medical Record Data Analysis and Prediction of Stroke Disease Using Explainable Artificial Intelligence (XAI). *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. **7**(4). https://doi.org/10.22219/kinetik.v7i4.1535.

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, **47**(1): 31–39. https://doi.org/10.17849/INSM-47-01-31-39.1

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* **3**(22):41-46.

Zihni, E., McGarry, B., & Kelleher, J. (2022). Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients. *Exon Publications*, 73-90. Retrieved January 26, 2023, from https://exonpublications.com/index.php/exon/article/view/explainable-decisions/941

Zhang, Z., & Han, Y. (2020). Detection of Ovarian Tumors in Obstetric Ultrasound Imaging Using Logistic Regression Classifier With an Advanced Machine Learning Approach. *IEEE Access*, **8**: 44999–45008. https://doi.org/10.1109/ACCESS.2020.2977962