Evaluation of a Stroke Prediction System: Performance, Transparency, Fairness, and Beyond

Jialing Li
jl9716@nyu.edu

Yuheng Yang
yy2597@nyu.edu

1 Background

According to the World Health Organization, stroke accounts for approximately 11% of all deaths globally, and many survivors face long-term disabilities. Accurate stroke early detection and medical support are crucial in reducing the risk of fatal outcomes. With the rise of AI, automated diagnosing systems have been employed by increasing numbers of healthcare professionals and institutions in medical imaging, epidemiology forecasting, and so on. In the case of stroke detection, an ADS predicting whether or not a stroke will occur for a patient analyzes vast amounts of Electronic Health Records (EHR) to identify patterns, potentially helping healthcare providers make more accurate decisions with lower costs compared to human doctors. However, responsible implementation of such a system requires us to carefully understand the transparency of algorithmic decisions, the potential bias against certain subgroups of patients, and the corresponding social implications. We intend to perform a technical audit of a stroke prediction diagnosis system, and carefully study the aspects mentioned above.

As an implementation of a stroke prediction system, Kaggle user Saimon Dahal presents his code implementation[1] based on his designed data processing pipelines and proposes a model based on Random Forest. The goal of the system is to accurately predict which patients are going to have a stroke in the future. Beyond this general purpose, this task involves a trade-off between two parties of stakeholders: hospitals and healthcare providers, to whom good precision is desirable as it helps facilitate the efficient allocation of medical resources; and (potential) patients, to whom good recall is important because it means the model is capable of capturing a high-risk individual in such a costly event. This precision-recall trade-off arises because a model with high precision tends to predict positive instances only when it is very confident; however, this may also cause it to miss some actual positive instances, leading to lower recall, and vice versa. We will further discuss this concern when implementing tests and audits to the model in the later section.

2 Input and Output

The dataset[2] with which Saimon Dahal trains the algorithm comes from Kaggle and the original owner of the dataset is HealthData.gov from the United States Government Department of Health & Human Services. Each record in the dataset is derived from the Electronic Health Records (EHR) of an individual. Altogether, the dataset contains 5110 entries and 12 columns, of which the information is summarized in table 1. We can see the dataset columns are well-defined and the only feature with missing values is bmi with a relatively low missing rate (3.9%).

Data Profiling Summary

Column	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Data Type	int64	object	float64	int64	int64	object	object	object	float64	float64	object	int64
Completeness	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete	201 Missing	Complete	Complete
Explanation	id	gender	age	whether hypertension	whether heart disease	if ever married	work type	area of residence	glucose level	BMI	whether smokes	whether stroke

Table 1: Data Profiling Table

To understand the distribution of the values in each column and how they relate to the outcome of whether an instance will have a stroke, we further plotted bar charts for categorical columns and used kernel density estimation to show the distribution of numerical columns. The results are shown in Fig. 1, 2, and 3. This preliminary

visualization shows that the dataset is imbalanced with fewer instances associated with a positive stroke outcome (which corresponds to the reality that stroke is not that common in all patient populations). To effectively train a model with the imbalanced dataset, Saimon Dahal implements SMOTE to upsample the minority class which gives a good validation performance. We will further discuss the method in the Implementation and Validation section and examine its legitimacy in terms of fairness in the Outcome section. Using the KDE plots, we find there exists a difference in the shapes of the distributions between the stroke individuals and non-stroke individuals. Approximately, higher ages, higher average glucose levels, and higher BMI are all associated with higher probabilities of stroke. We anticipate these numerical features will be important determinants in the model. The features do not show a strong correlation.

The output of the system is a class label, which indicates the model's estimation result of whether a stroke will occur to an individual in the future. If the model gives an output of 1, it predicts the instance will have a stroke, and 0 otherwise.

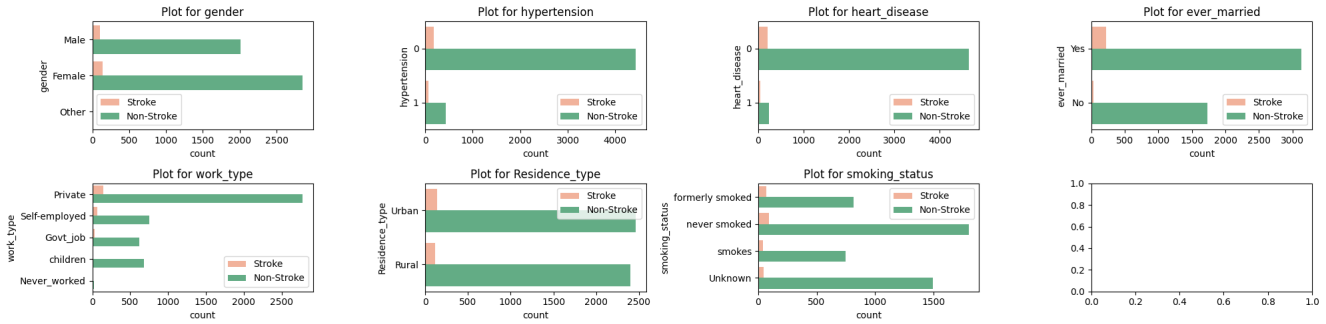

Figure 1: Categorical Column Distributions

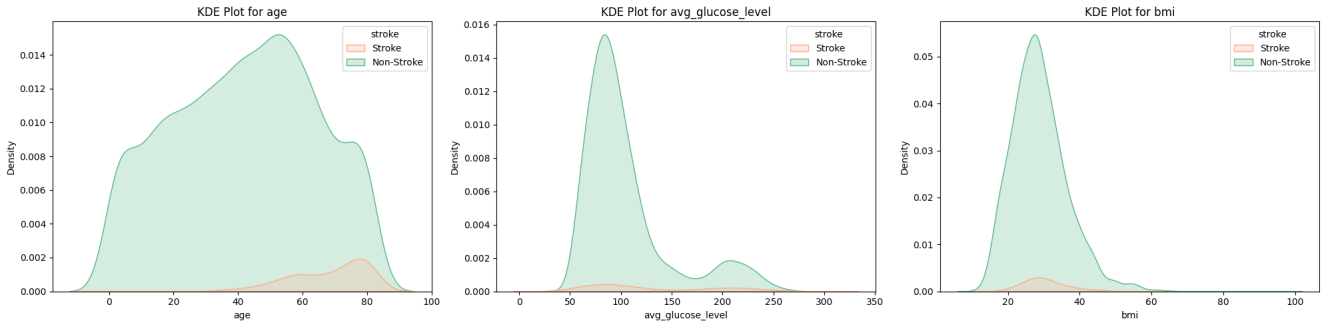

Figure 2: Numerical Column Distributions

3 Implementation and Validation

The provided code in the original ADS designs a pipeline to process and model the dataset to predict whether a stroke will occur to an individual recorded in the EHR database.

3.1 Data Cleaning and Pre-processing

The author starts with an Exploratory Data Analysis that visualizes the relationship between various features. Then, the author cleans the data by removing rows with missing values and separating the features and the target variable 'stroke'. The data is split into training and test sets, with numerical features standardized using StandardScaler and complex categorical features encoded using OneHotEncoder. Unique mappings are applied to convert other categorical variables like 'ever-married' to the indicator format for further model training.

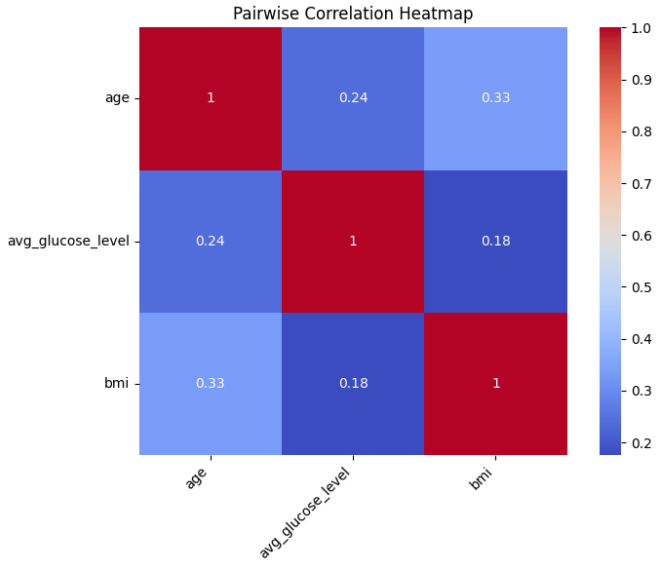

Figure 3: Correlations

3.2 The Implementation of the System

As discussed before, there is an imbalance of the target variable in the dataset, with negative labels (no stroke occurs) significantly more than positive labels. This imbalance could influence the predictive modeling, as the over-representation of the non-stroke category may skew the model's ability to generalize to the less-represented data. To address the class imbalance issue, the original ADS employs SMOTE (Synthetic Minority Over-sampling Technique) on the training set to generate synthetic samples for the minority class (patients identified with stroke), which derives a 50:50 balanced training sample set. With the author's random setting, we present the data distribution of this balanced dataset in Fig. 4 and Fig. 5.

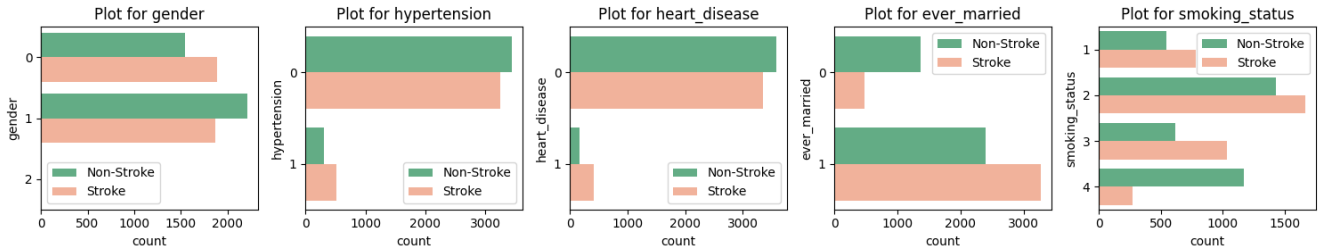

Figure 4: Categorical Column Distributions (Balanced)

Training on this balanced training set, the author experimented with logistic regression, random forest classifier, support vector classifier, and k-nearest neighbor classifier, all with default settings. The author concluded that the random forest classifier has the best performance and performed hyperparameter tuning with 5-fold cross-validation. Finally, the author proposed a random forest classifier model with {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}. Notably, when choosing which model to further tune on, the author used the test set to evaluate the accuracy. We deem this inappropriate as it may introduce data leakage and performance inflation. However, since the author proposes the random forest method, we decided to stick with this final selection in our further explorations.

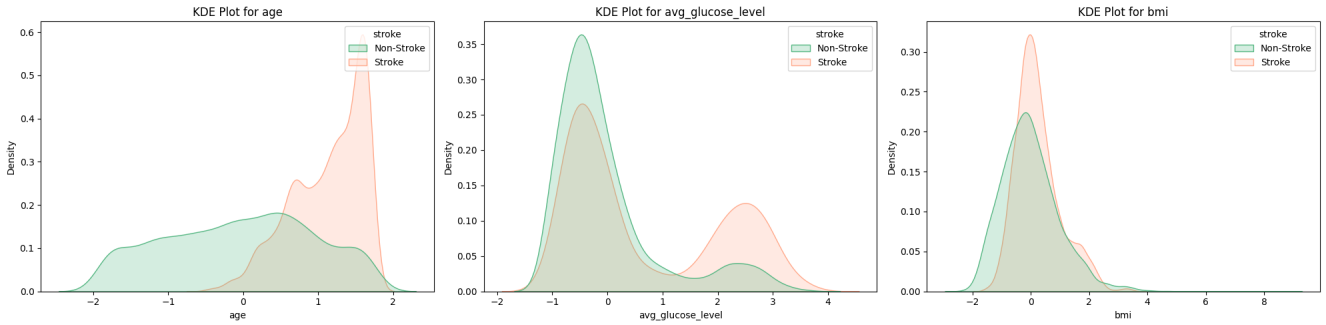

Figure 5: Numerical Column Distributions (Balanced)

3.3 Validation of the ADS

The author uses the accuracy rate as the main metric of optimization. In the 5-fold cross-validation, each fold contains sample points with approximately balanced target variables because they are randomly drawn from a balanced dataset upsampled by SMOTE. With these (balanced) folds, the author's proposed model achieves a 0.96 average accuracy, which significantly outperforms the 0.50 random guess baseline. The author deems this model successful, but as will be shown later in the following sections, we find this model inappropriate in the real-world setting where data imbalance persists in the EHR records, and it also introduces significant biases across gender subgroups.

4 Outcomes

In this section, we demonstrate how and why the ADS proposed by the author is flawed. In specific, we found the ADS, though looks promising in the cross-validation phase, does not perform well with a real-world setting unseen data, and it creates disparities in the performance between males and females and others. In fact, within the stakeholder group of patients, the ADS's performance deteriorates when facing individuals of gender being females and others. To ensure the robustness of our arguments, we will perform our audit on data splits generated by 10 random seeds and make box plots to examine the distribution of our selected key metrics on the unseen test sets. All other procedures remain the same as proposed by the author. We will also discuss the transparency of the ADS by studying how it makes decisions, and discuss other concerns at the end.

4.1 ADS Performance

As discussed in the Introduction section, the stakeholders of our problem involve two parties: healthcare providers who demand a conclusive and reliable model to allocate the treatments efficiently and avoid the waste of medical resources; and the (potential) patients who demand accurate identification of those who are really going to experience a stroke, reducing misidentification and disastrous outcomes. This tension highlights two selections of the performance metric respectively: precision, which evaluates the proportion of predicted positive cases being true positive; and recall, which evaluates the proportion of true positive cases successfully identified. They address the concern of each group of stakeholders. As well as these, we also compute the accuracy of the model as a highly accurate model benefits everyone. The result of our experiment run on test sets of different data splits is shown in Fig. 6.

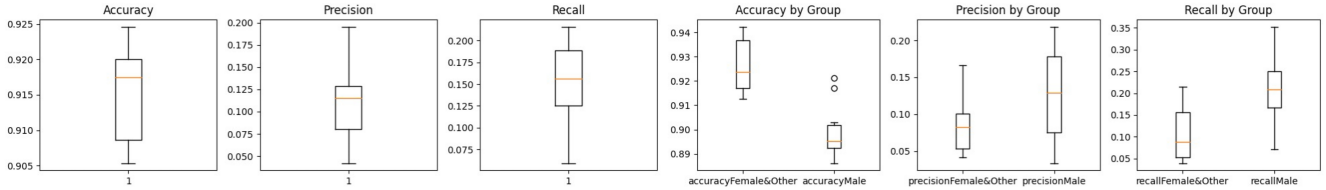

Figure 6: Performance Metrics

One thing we immediately notice is that although the model is capable of producing high accuracy in both validations (discussed in 3.3) and test sets, the test result cannot be seen as successful because the test sets are not balanced by SMOTE, that is, they reflect the true distribution of strokes in the population. The stroke by itself is a relatively rare event: in the original dataset, the negative labels account for 91.9% of the records. If we consider this reflects the true possibility and test sets are drawn by complete random (this is likely to be true given we performed 10 random splits), then predicting with the majority class should give us an accuracy about 91-92%. Shown in Fig. 6, the proposed model just performs as well as this benchmark, meaning that the model is just as good as the majority class vote method.

Apart from that, we found the model has a low overall precision (around 11%), and at the same time, an unsatisfying recall (around 16%). This means the proposed model does not perform well in either helping healthcare providers accurately allocate medical resources, or correctly identifying those potential patients with real needs, especially when the data reflects real settings.

Extending the calculation to by-group performance, we found that the performance metrics on males are significantly better than on females and others, characterized by higher precision (~0.13 vs. ~0.09) and higher recall (~0.23 vs. ~0.09). This shows a deteriorated model performance with females and other individuals. In particular, with a much lower recall, truly risky females and others are more likely than males to be overlooked by the ADS. This will be explained further in the next section, where we unpack this disparity using more fairness metrics.

4.2 ADS Fairness

4.2.1 A Linkage to EO

By the process itself, the distribution of medical resources to those truly in need resembles the college admission problem and the loan approval issue, as they are all problems regarding *the allocation of limited resources*. This hints us to evaluate the ADS and apply fairness metrics through the lens of theories of equal opportunities doctrines. In our scenario, the goal of the implementation is to successfully predict the *future* possibility of stroke. We also hope that the system will treat each gender subgroup fairly, in the sense that it does not systematically overestimate or underestimate the stroke occurrence versus the truth in the future in one subgroup than the other subgroup. It requires that the error made by the ADS does not track group membership. This forward-facing problem and its considerations fit into the idea of Formal-Plus EO. As explained by the 2022 work of Khan et al.[3], the fairness metrics, error rate balance and equalized odds, align with the idea of Formal-Plus EO. Given we set our scenario in the perspectives of the Formal-Plus EO, we decided to use these metrics as our selection of fairness metrics. According to Chouldechova's 2017 work [4], error rate balance requires false positive and false negative error rates to be equal across the groups. To more closely examine the ADS, we follow Chouldechova's idea and unpack the error rate balance metric into calculating the false positive rate (FPR) and false negative rate (FNR) for each gender subgroup and make comparisons. We will record the FPR difference and FNR difference across gender groups at each run and they give us the absolute difference, which allows us to see if any disparity exists. As well as that, we also calculate FPR of males minus FPR of females and other (denoted by FPR M-F&O); and FNR of males minus FNR of females and other (denoted by FNR M-F&O), which enables us to evaluate the direction of the bias, if any. These metrics closely align with the real-world implications as they measure if the ADS imposes disparities on falsely sending people into hospitals, or more seriously, neglecting the real risk of stroke. We will also calculate equalized odds at each run as it measures how the model performs (un)equally across different subgroups.

4.2.2 Bias in SMOTE

In addition, we found auditing the data processing pipeline necessary because the author applied SMOTE to upsample the minority class and used the created dataset as the training set. This may alter the true data distribution because the SMOTE technique upsamples the minority class by a distance-based method. It first calculates the Euclidean distance between minority instances. After that, it chooses the k-nearest neighbors of a given instance and randomly connects one of them to the given instance in the feature space. A new sample is thus generated by a convex combination between the two instances. It makes SMOTE susceptible to biases caused by distinctive features dominating a subgroup. For example, if stroke is associated with hypertension and hypertension is more common in males, then very likely the newly created samples are still males. This may over-represent one group, and/or exacerbate the pre-existing bias in the data. In fact, we observe it does happen in the author's solution. In Fig. 1, we see the likelihood of stroke is more or less similar within males and females; however, in Fig. 4 (after SMOTE), we see that the number of instances with stroke in males even exceeds that of instances without stroke in

males, while for females and others, the non-stroke instances still account for the larger share. This shows that the males are over-represented in this data pipeline, possibly leading to inflated performance for males. We suspect this happens systematically. To monitor the SMOTE, we define our own fairness metric, SMOTE Inflation Factor (SIF), as calculated by

$$SIF = \frac{\frac{\text{proportion of stroke in male after SMOTE}}{\text{proportion of stroke in females and other after SMOTE}}}{\frac{\text{proportion of stroke in male before SMOTE}}{\text{proportion of stroke in females and other before SMOTE}}}$$

The calculation formula is self-explanatory: by recording the SIF at each run, we can monitor if the positive instances are upsampled in its original proportion across genders and if any subgroup is over-represented in the SMOTE process.

4.2.3 Results

Having defined our fairness metrics of concern, we run them on the same 10 different splits as in 4.1. The results are recorded in Fig. 7.

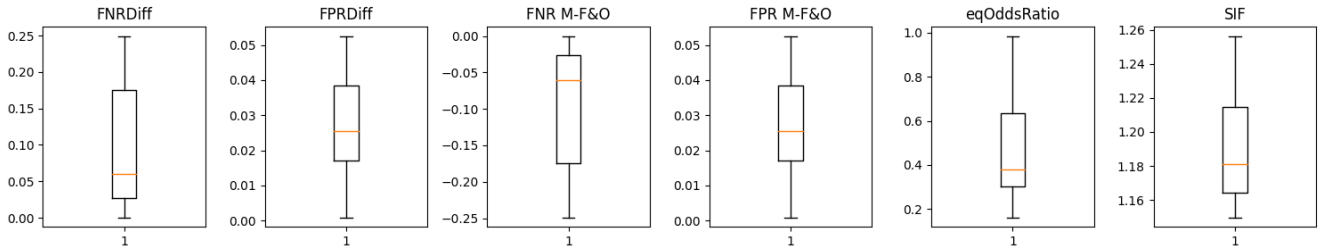

Figure 7: Fairness Metrics

As we can see, the FNR and FPR differences are always above 0, and the variation of the FNR difference is especially large. In addition, the equalized odds ratio significantly differs from 1, with a median of 0.4, and primarily lies from around 0.3 to around 0.7. This evidence shows that the ADS systematically generates disparities in errors across different gender groups, and it does not satisfy the error rate balance or equalized odds requirements of the Formal-Plus EO. In a closer look, we notice that the FNR M-F&O is always smaller than zero; and FPR M-F&O is always larger than zero. Notably, the FNR M-F&O metric can sometimes reach as high as -0.15 to -0.20. These results show that the ADS tends to generate more false negative outcomes for the females and other group. This is very harmful to the patients stakeholders because it means that the model is systematically missing more risky individuals in females and other compared to males. When we focus on the SIF metric, we notice it is always larger than 1. It means that the data processing pipeline is constantly amplifying the proportion male samples in the positive class than in the original dataset, distorting the true data distribution. It serves as a good piece of evidence showing the SMOTE process is inflating the representativeness of males. It may introduce or exacerbate the system's biased recognition that males are more associated with stroke than females and other are, which relates to the result of FNR M-F&O and FPR M-F&O that there are fewer false negatives and more false positives in males.

To summarize, through the experiments of the fairness metrics, we found that the ADS shows systematic disparities in the error rates across different gender groups. From the standing points of patients, the model is more likely to underestimate the risk with the females and other group than males group. The data processing pipeline may worsen the situation. The ADS cannot be concluded as a fair system.

4.3 ADS Interpretability

To further evaluate ADS performance and enhance the interpretability of the model as well as its transparency, we implemented SHapley Additive exPlanations (SHAP), applying feature importance scoring and force plots to gain insights into which variables most significantly impact the model's decisions. This is critical in understanding potential biases and the factors that contribute most to the prediction of a stroke.

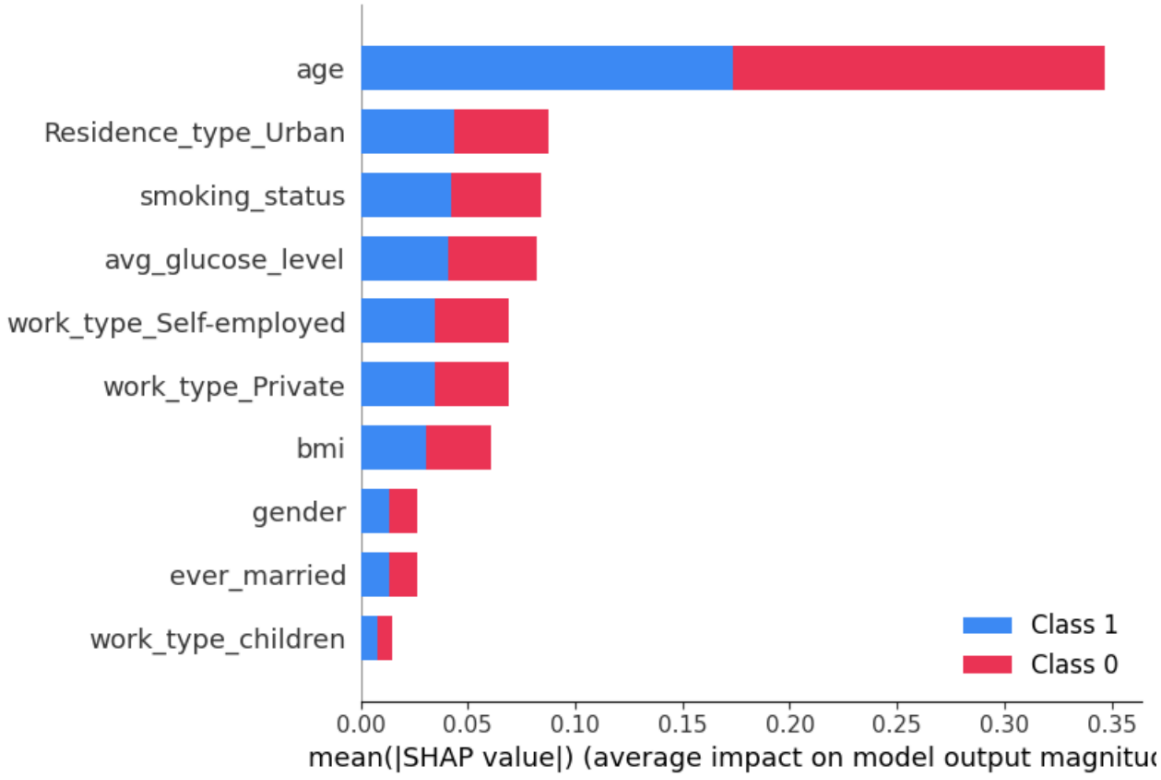

Figure 8: Feature Importance

4.3.1 Feature Importance: Global Interetability

Based on the feature importance plot produced from the model, we found that age is the most significant factor in determining whether a patient will have a stroke or not, which has an average impact on model output magnitude significantly larger than other features. Following age, residence type, smoking status, and glucose levels are the next most influential factors. Overall, these results align well with our expectations and make sense given the known risk factors for stroke. However, to facilitate a more transparent understanding and to evaluate whether the model can be trusted of not, we also need to examine closer to individual prediction processes.

4.3.2 Forceplot: Local Interpretability

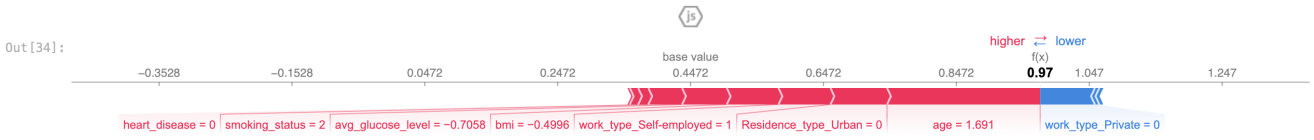

Figure 9: local interpretability

We then advanced to local interpretability, utilizing SHAP force plots to provide detailed insights into individual predictions. These plots reveal how different feature values, depicted in pink for increasing the possibility of predicting as having stroke and blue vice versa, influence the probability of predicting a stroke. The magnitude of each bar represents the effect of that feature, and the cumulative impact of all feature SHAP values explains the deviation of the model's prediction from the baseline.

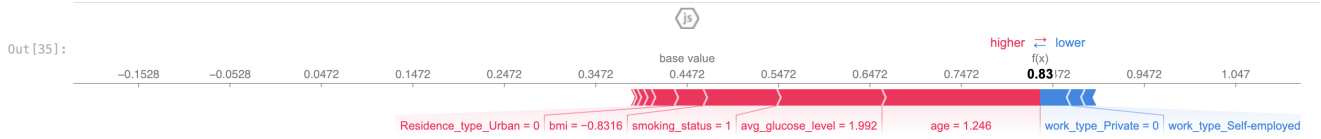

Figure 10: local interpretability

Figure 11: local interpretability

To demonstrate the practical application of these interpretative tools, we randomly selected three instances where the model's predictions align with the actual conditions.

In the first instance, as shown in Figure 9, the model assigned a score of 0.97 for a stroke, compared to the base value of 0.4472. The major factors elevating this prediction were the patient's age and its residence type as urban residents, both indicative of increased stroke risk. The second case depicted a stroke probability of 0.83. Here, the significant contributors were the patient's age (standardized value of 1.246) and average glucose level (standardized value of 1.992), aligning with established stroke risk factors related to advanced age and elevated glucose levels. In the third instance, the model predicted a very low probability of a stroke at 0.03, compared to the base value of 0.4472, indicating a prediction of non-stroke. The major contributing factors to this outcome are a low feature value in age (standardized value of 0.0002361) and a non self-employed work type, which both significantly lowered the prediction probability. These factors suggest that younger age and certain occupational profiles may be associated with a reduced risk of stroke. It highlights the nuanced influence and interactive effects of demographic and socioeconomic elements in the model's assessments.

These case studies, illustrated through SHAP force plots, demonstrate how individual feature contributions lead to specific prediction outcomes, aligning with the global feature importance results. Overall, we notice that the model makes sense as how it utilizes features to score up or down the risk of a stroke aligns with our consensus most of the time on this medical issue (e.g., older age and higher glucose level are associated with higher risks). However, we noticed that there is a potential bias in how smoking status is encoded in the model; values are assigned as 'smokes': 1, 'never smoked': 2, 'formerly smoked': 3, 'Unknown': 4. This ordinal encoding might lead the model to misconstrue the numerical value as a scale of severity or frequency of smoking, which could distort the predictive accuracy and model interpretability. This is shown in the Fig. 9 and Fig. 10, where we notice that smoking_status of 1 and 2 both contribute to higher risks. While 1 indicates smoking, 2 represents never smoked. In this specific feature, medical practitioners can hardly trust the evaluation results given by the ADS. The correct approach would typically involve using one-hot encoding to prevent any ordinal assumptions by the model about the categories.

Although the individual case analysis suggests that the predictions are consistent with most known risk factors, the misleading data processing applied on the smoke_status feature introduces confusion and harms the level of trust. Considering the ADS's disparities in different gender groups as we mentioned above, we do not recommend practitioners rely on this model.

5 Summary

In conclusion, we find the ADS implemented by Saimon Dahal inappropriate for real-case applications. Deploying this ADS in the public sector of hospital admission and disease early detection is highly problematic because when facing real-world, unseen data, the ADS has an undesirable performance, and the situation is worsened by unfair prediction results it generates for different gender groups.

From the perspective of performance, we recognized that two groups of important stakeholders are healthcare providers, who benefit from a conclusive model and aim at allocating medical resources efficiently; and (potential) patients, who most benefit from a system that correctly identifies all positive cases. They respectively require high precision and high recall. However, through our experiments in section 4.1, we found that the ADS has a stable low precision (around 11%) and a stable low recall (around 16%) in the test set. The author did not correctly identify these stakeholders and their needs and instead chose to optimize towards a high accuracy. However, as explained in detail in 4.1, the occurrence of stroke is not that common in the real world. Assuming the original dataset reflects the true target variable distribution, the ADS is just performing as well as blindly predicting the majority class. Our experiments on these performance metrics show that the model does a poor job at its stated goal, and fully relying on its prediction results causes inefficiencies and dangers to both hospitals and patients. Therefore, we conclude the implementation does not address the accuracy or robust concern of a properly designed ADS in the medical sector.

In further exploration, we find that the data may be a source of the problem. The dataset with which Saimon trains the model, though recorded with good quality, is highly imbalanced in the target variable. This may reflect the real situation, however, it also causes serious troubles for algorithms to comprehensively understand the pattern of the data distribution within each target class, as the positive samples are very underrepresented. To fix this, the author applied SMOTE. However, as we have examined in section 4.2, this method may have exacerbated the bias across genders because it may overrepresent or underrepresent a group disproportionally. We are also concerned that the few positive samples will limit the ADS's ability to generalize to unseen patient cases. Therefore, we deem the dataset inappropriate and insufficient for the given task.

We also analyzed our concern regarding the fairness of the ADS through the lens of Equal Opportunities. As explained in 4.2, we found the Formal-Plus EO to be good theoretical framework of evaluating this ADS because it is a forward-facing problem that concerns the error rate within each gender group. As well as defining the SIF fairness metric and finding that the data processing pipeline may be a source of technical bias as discussed in the paragraph above, we also evaluated the ADS against error rate balance and equalized odds. In specific, a low FPR may benefit healthcare providers as it avoids resource waste, and a low FNR benefits patients as it reduces the risk of being harmed by stroke. In experiments, we found that the model's error does track group membership and thus becomes incompatible with the Formal-Plus EO idea. Specifically, we find the ADS systematically underestimates the likelihood of stroke for the female and other group, characterized by a stably higher FNR. We found the ADS violates the fairness concern as it introduces more risks to this group, especially at such as high-stake task of disease detection.

By further evaluating the ADS and assessing its interpretability as well as transparency with SHAP techniques, we have gained insights into how the model makes its predictions. The use of both global and local interpretability tools, such as feature importance plots and SHAP force plots, has provided a clear and detailed view of the factors that most significantly influence outcomes, and it also allowed us to spot another feature processing practice that damages the trust of this ADS. We conclude that the sub-optimal feature encoding method and the disparities the ADS creates seriously undermine the trustworthiness of this model.

Having systematically explored the ADS and the dataset, we propose several recommendations to mitigate the unfairness and treat the current process. Firstly, we advise that this type of model should be trained on a more balanced dataset that represents each label sufficiently. At a minimum, it should involve a more diverse feature distribution for each class, so that the model can learn the pattern more comprehensively and generalize better. Secondly, we suggest an evaluation before implementing the ADS, and the model builder should carefully choose which metric to optimize on. For instance, if the author chose to optimize FPR, FNR, or F-1, the ADS may have better actual performance when addressing the concern of each stakeholder. Thirdly, when data filtering/aggregating/generating process occur in the pipeline, we recommend defining a customized metric (e.g., SIF in our case) to track the distribution change regarding the sensitive features to ensure the process does not incur technical bias that underrepresent any subgroups. Fourthly, we recommend that ADS builders should use ordinal encoding very carefully because it may lead to false numerical relationships between categories. Instead, builders should consider using strict one-hot encoding when facing these features. Last but not least, we recommend timely evaluating the fairness metrics across groups and should there be any biases, the ADS builder should consider involving post-processing (e.g., customizing threshold for each subgroup) to minimize the disparity.

References

[1] Saimon Dahal. 2024. Stroke Prediction [Random Forest 91%+ Accuracy]. *Kaggle*.
https://www.kaggle.com/code/saimondahal/stroke-prediction-random-forest-91-accuracy

[2] Stroke Prediction Dataset. 2021. *Kaggle*.
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

[3] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines.
https://doi.org/10.48550/arXiv.2207.02912

[4] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5:2, 153–163, DOI: 10.1089/big.2016.0047.