Stroke Risk Detection with ML

Project Members: Pragnya Velivela, Reeda Huda, Amelia Thomas, Rhea Iyer, Harinishree Sathu

Introduction | Problem Definition | Methods | Results/Discussion | References | Contribution Table/Gantt Chart

Proposal Report

Introduction

Machine learning has become increasingly valuable in healthcare, enabling personalized predictions and improving patient outcomes by considering multiple aspects of patient data while coming to a final diagnosis. Models discussed in Predictive modelling and identification of key risk factors for stroke using machine learning, such as decision trees, random forests, SVMs, and neural networks—offer insight into the data processing and feature extraction needed for stroke prediction [2].

https://www.kaggle.com/datasets/gustavojota/samsung-heart-rate-fit2

In this study, we aim to predict stroke risk using various demographic and health-related factors. This Kaggle dataset has information about 5110 patients at risk for stroke and includes attributes such as age, hypertension, heart disease, and lifestyle indicators (like smoking status), as well as physiological measures like BMI. https://www.nature.com/articles/s41598-024-61665-4

Problem Definition

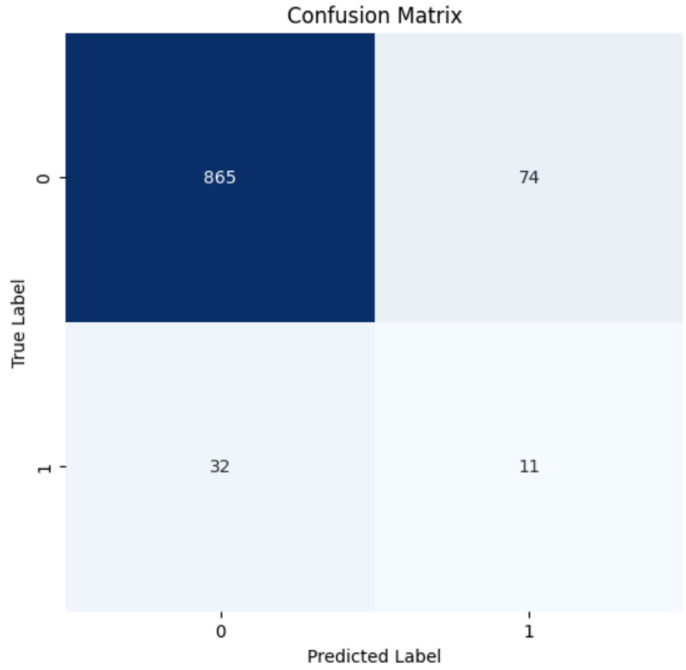
Cardiovascular diseases, such as strokes, are influenced by various demographic and health factors. Predictive models that assess stroke risk based on these factors could significantly improve early detection and prevention efforts. This project aims to leverage demographic and lifestyle data—such as age, hypertension, heart disease, BMI, and smoking status—alongside machine learning algorithms to predict stroke risk. By analyzing this data, our model seeks to support healthcare providers in effectively identifying at-risk individuals.

Methods

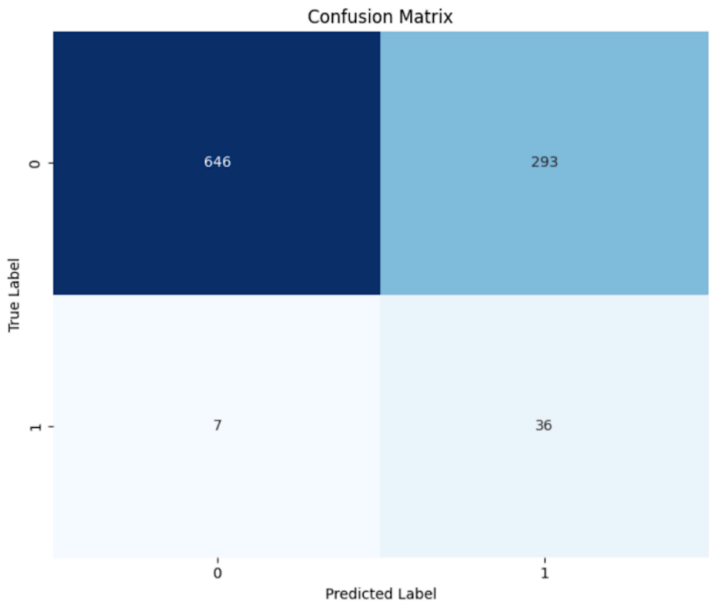
In terms of data preprocessing, we utilized scikit-learn's LabelEncoder to convert the categorical variables in our dataset(ex. gender, work type, marital status) to either 0 or 1, which allows the model to interpret the data accurately. We then applied scikit-learn's StandardScaler to normalize numerical columns, excluding discrete variables such as stroke, hypertension, and heart disease, which reduces bias amongst these features. Then, we incorporated a check for any rows with missing values or NaN to ensure that didn't affect the results of the model and visualized each of the features with matplotlib, which allows us to see if there were any further data preprocessing steps that needed to be taken. Finally, we split our data into 80% training and 20% testing, which we saved into respective csv files for use in different notebooks for each machine learning model.

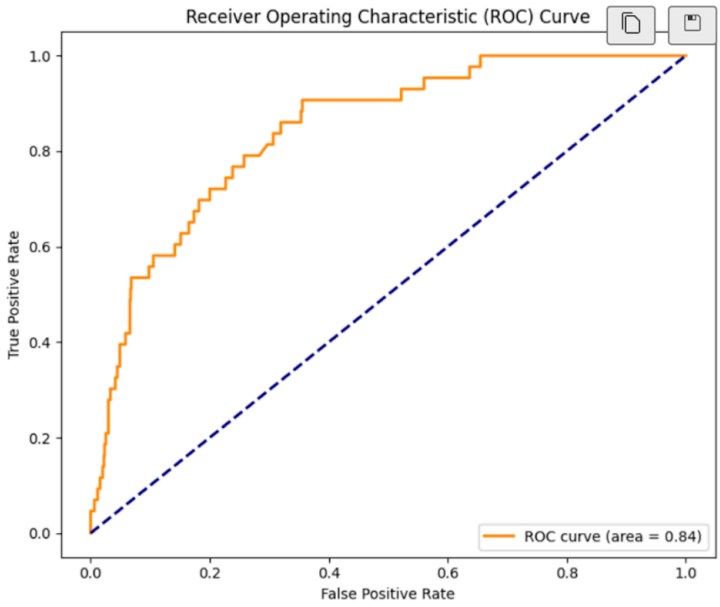
Results/Discussion

The Random Forest model trained with SMOTE-balanced data shows uneven performance. It performs well for the validation test at predicting non-stroke cases, with a high precision (96%), recall (91%), and an F1-score of 0.93. However, when it comes to stroke cases, the model struggles significantly, with a precision of only 9% and recall of 20%, resulting in a very low F1-score of 0.12. On the test set, the trend is similar: the model achieves strong performance for non-stroke cases (F1-score of 0.94), but for stroke cases, it only manages a precision of 12%, recall of 26%, and an F1-score of 0.16. This indicates that, despite balancing the training data, the model is still heavily biased toward the majority class. When trained on unbalanced data, the model predicts all cases as non-stroke, failing entirely to identify stroke cases. These results highlight the need for further refinement, such as adjusting class weights, enhancing the dataset, or trying different algorithms to better capture patterns for stroke cases.

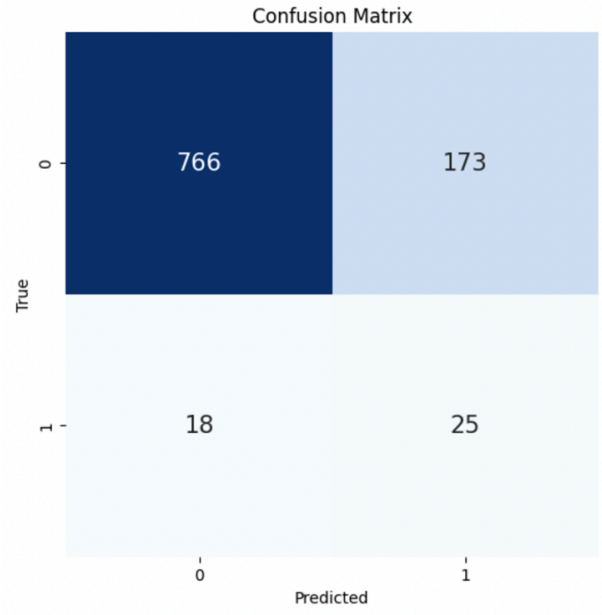


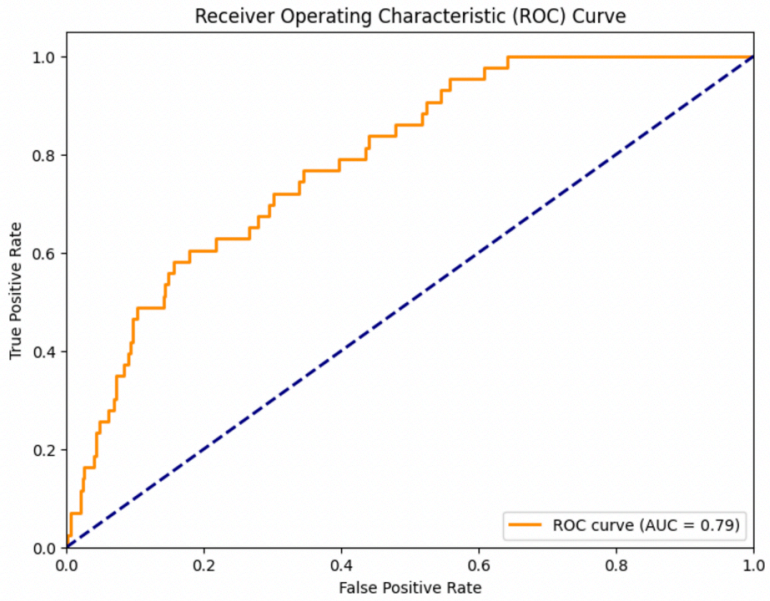
The SVM model achieved an overall accuracy of 69.45%, providing a general measure of how often the model can predict correctly. The precision for the non-stroke class is extremely high at 99%, meanwhile the precision for the stroke class is very low at 11%. This can be attributed to the significant class imbalance in the dataset. This low precision results in a high number of false positives, where non-stroke cases are incorrectly predicted as strokes. For the recall metric i.e., the model's ability to identify true cases of each class, the model achieves a recall of 69%. For stroke cases, the recall is much higher at 84%, reflecting the model's ability to correctly identify the majority of actual stroke cases. The F-1 score combines precision and recall, giving an overall measure of the model's performance for each class. For non-stroke cases, the F1-score is 0.81, indicating a good balance between precision and recall. However, for stroke cases, the F1-score is much lower at 0.19, showing that the model struggles to perform well for the minority class. In summary, while the SVM model demonstrates strong performance in identifying non-stroke cases and achieves high recall for stroke cases (84%), its low precision for strokes (11%) significantly limits its effectiveness. The high number of false positives leads to a poor F1-score for stroke cases (0.19), indicating that the model struggles to balance precision and recall for the minority class. These results highlight the need for further improvements, such as addressing class imbalance and refining the model to enhance precision, to make it more reliable.





Lastly, the RNN model demonstrates a strong performance in predicting non-stroke cases. However, it does face challenges in effectively identifying stroke cases. It receives an overall accuracy of 91%, but there is a significant disparity in how well it can handle the two (stroke and non-stroke) classes. For non-stroke cases, the model performs exceptionally well. With a precision of 98%, almost all predictions for non-stroke cases are accurate. Additionally, the model achieves a recall of 82%, correctly identifying the majority of actual non-stroke cases. These metrics combine to produce a strong F1-score of 0.89, reflecting a good balance between precision and recall. However, for stroke cases, while the model achieves a recall of 58%, meaning it correctly identifies more than half of the actual stroke cases, its precision is only 13%. This indicates that the majority of cases predicted as strokes are actually non-stroke cases, leading to a high number of false positives. The resulting F1-score for stroke cases is just 0.21. In conclusion, the RNN model performs well for non-stroke predictions, demonstrating high precision, recall, and F1-score for the majority class. However, its poor performance on stroke cases, with low precision and F1-score, limits its performance.





When comparing the three models, the Random Forest stands out for its overall strong performance in predicting non-stroke cases, delivering consistently high precision, recall, and F1-scores. However, like the SVM model, it struggles to accurately identify stroke cases. The SVM model shows promise with a high stroke recall of 84%, meaning it identifies most stroke cases, but its extremely low precision of 11% leads to a high number of false positives. This significantly undermines its reliability for practical use. Meanwhile, the RNN model strikes a balance, offering better stroke recall than Random Forest, but with slightly lower performance in non-stroke predictions. However, its low precision across the board for stroke cases highlights an ongoing challenge. Ultimately, the best model depends on the specific priorities of the application. If the goal is to minimize missed stroke cases, the

higher recall of the SVM and RNN models makes them more suitable candidates. However, their low precision suggests the need for further refinement to reduce false positives. On the other hand, the Random Forest model is more balanced overall, excelling at non-stroke predictions but falling short in identifying strokes, which limits its suitability in critical medical applications where detecting strokes is paramount.

References

1. K. Feldman, R. G. Duncan, A. Nguyen, G. Cook-Wiens, Y. Elad, T. Nuckols, and J. M. Pevnick, "Will Apple devices' passive atrial fibrillation detection prevent strokes? Estimating the proportion of high-risk actionable patients with real-world user data," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 6, pp. 1040-1049, Jun. 2022.
2. N. Mohammadian Rad and E. Marchiori, "Chapter 9 - Machine learning for healthcare using wearable sensors," in *Digital Health*, A. Godfrey and S. Stuart, Eds. Academic Press, 2021, pp. 137-149. doi: 10.1016/B978-0-12-818914-6.00007-7.
3. P. Weerakody, K. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *ScienceDirect*, vol. 453, pp. 773-788, 2021. doi: 10.1016/j.neucom.2021.02.046.

Contribution Table and Gantt Chart

Gantt Chart link: Google Sheets Link

Contribution Table

Name	Proposal Contributions
Rhea Iyer	GitHub Pages, Contribution Table, Updated Gantt Chart, Video
Harini Sathu	Fixed SVM Model, Video
Amelia Thomas	RNN Model, SVM Model, Video, Created Gantt Chart
Reeda Huda	Random Forest Model, Analysis + Comparison, Video
Pragnya Velivela	Random Forest Model, Next Steps, Video