Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Development of Algorithms for Early Detection and Prevention of Diabetes in Patients with Prediabetes Living in Rural or Underserved Population AND Prevention of Ischemic Stroke in Patients with Atrial Fibrillation (AF)
Principal Investigator:	Arinze Nkemdirim Okere
Prepared by (IIAI):	Yanan Liu
Other investigators:	Joshua Carrizales, Alperen Duyan
Date:	

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	2/27/2025
Funding agency	The American Heart Association (AHA)
Grant mechanism	
If no … Why not? What went wrong?	

Brief summary of accomplished results:

This study investigates the use of machine learning to predict 5-year stroke risk in patients newly diagnosed with atrial fibrillation (AF). Using a dataset of 11,541 patients, we trained a random forest classifier on clinical and laboratory data collected within one month of the AF diagnosis. Predictors included ECG parameters, lab values, vital signs, demographics, smoking status, and medication history. The model achieved a test accuracy of 67%, with a sensitivity of 0.624, specificity of 0.680, precision of 0.293, and an F1 score of 0.398 for stroke and test accuracy of 66.2%, with a sensitivity of 0.615, specificity of 0.674, precision of 0.311, and an F1 score of 0.413 for diabetes. These results highlight the potential of machine learning in clinical risk prediction, though further refinement is needed. Future work will focus on narrowing the feature set to the most predictive variables and exploring advanced modeling techniques to improve performance and clinical relevance.

Research report:

Aims (provided by PI):

Aim 1: To identify early predictors for ischemic stroke and develop a predictive algorithm for ischemic stroke.

Aim 2: To compare and determine the best ablation procedure most likely to prevent stroke recurrence.

Aim 3: To develop an algorithm that identifies key factors influencing the progression from prediabetes to diabetes using a neural network approach.

Aim 4: To create a deep learning-based algorithm for the early screening of patients at high risk of transitioning from prediabetes to diabetes.

Due to time limitation and multiple aims, we focused on predict stroke risk/diabetes.

Data:

This study used a retrospective cohort design based on electronic health records (EHR) to investigate the five-year stroke risk in patients newly diagnosed with atrial fibrillation (AF). Patients were included if they had a first-time AF diagnosis (ICD-10: I48.91) between 2013 and 2019 and at least five years of follow-up. Those with a prior stroke within three years, or under the age of 18, were excluded. The index date was defined as the date of AF diagnosis, and the primary outcome was the occurrence of a stroke within five years, identified using ICD-10 codes I63.9, Z86.73, and G45.9. Predictors were drawn from a wide range of clinical domains, including demographics, vital signs, laboratory results, ECG parameters, medication history, and smoking status. All features were selected based on their availability prior to or within one month after the AF diagnosis. Missing numerical values were imputed using population means to maintain model compatibility. The final dataset was randomly split into training (80%) and testing (20%) subsets. In a related analysis, a diabetes cohort was constructed by identifying patients with HbA1c values between 5.7 and 6.4 and tracking whether their HbA1c exceeded 6.5 within five years. Patients with at least two follow-up HbA1c results were retained, and a binary diabetes flag was assigned. This cohort was further enriched with vitals, labs, medications, and social history data, all aligned to the index date. Of the 4,157 patients in the diabetes cohort, 844 (20.3%) were flagged as diabetic. This structured approach to identifying and labeling diabetes status demonstrates the potential for applying similar methods to refine stroke prediction models by incorporating comorbid conditions.

AI/ML Approach:

To explore the feasibility of predicting stroke risk in patients diagnosed with atrial fibrillation (AF), a machine learning framework was developed using structured clinical data. The modeling pipeline was designed to accommodate a wide range of supervised learning algorithms, with the goal of identifying patterns in pre-diagnosis data that could be indicative of future stroke events. The dataset was preprocessed to ensure compatibility with machine learning workflows. This included encoding categorical variables using one-hot encoding, standardizing continuous features, and handling missing values through imputation or model-inherent strategies. Multiple machine learning models were implemented and evaluated to compare their performance and suitability for this clinical prediction task. These included ensemble methods such as Random Forests and Gradient Boosting (e.g., XGBoost), as well as classical algorithms like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees. Each model was trained using a consistent feature set and evaluated using standard classification metrics. The Random Forest algorithm was a central component of the modeling strategy due to its robustness to overfitting, ability to handle high-dimensional data, and interpretability through feature importance scores. Gradient boosting methods were also explored for their capacity to capture complex nonlinear relationships and interactions among features. SVM and KNN provided complementary perspectives, particularly in terms of decision boundary behavior and sensitivity to feature scaling.

All models were trained and validated using stratified cross-validation to ensure generalizability and to mitigate the effects of class imbalance.

Experimental methods, validation approach:

We treated stroke prediction and diabetes as binary classification problems. Our input features—they are explained in Data section—were preprocessed by one-hot encoding for categorical fields. Since we are using random forest classifier, we did not standardize numerical values as scales are not important in random forest algorithm. The target variable was a binary label (0 = no stroke, 1 = stroke); out of 11,537 total patient records, 16.8 % were positive for stroke. The dataset was split into an 80 % training set (n = 9229) and a held-out 20 % test set (n = 2308). Within the training set, 83% of the patients (7667 patients) were labeled as no stroke and 17% of the patients (1562 patients) were labeled as stroke. In the test set, 83.5% of the patients (1928 patients) were labeled as having no stroke and 16.5% of the patients (380 patients) were labeled as having stroke. From these numbers, it is obvious that we have a class imbalance, and it affected our results in a bad manner. Therefore, we used random down sampling which can be reproducible with random state = 42 on patients who are labeled as no stroke in our training set. We took 1562 non-stroke patients which is the same number of stroke patients. According to our parameter tuning, the best number for our number of trees is 130. We observed that accuracy, specificity, precision and F1 scores do not change significantly with the number of trees. However, there is a significant change in sensitivity with the changing number of trees. We treated diabetes prediction as a binary classification problem in the same way as stroke. Our input features (see Data section) were preprocessed by one-hot encoding for categorical fields; since we use a random forest classifier, we did not standardize numerical values. The target variable was a binary label (0 = no diabetes, 1 = diabetes); out of 4,157 total patient records, 844 (20.3 %) were positive for diabetes.

The dataset was split into an 80 % training set (n = 3,325) and a held-out 20 % test set (n = 832). Within the training set, 79.5 % of patients (2,642 patients) were labeled as no diabetes and 20.5 % (683 patients) as diabetes. In the test set, 81 % of patients (671 patients) were labeled as no diabetes and 19 % (161 patients) as diabetes. As with stroke, this class imbalance led us to apply random down-sampling (using random_state=42) on the majority class in the training set, selecting 683 non-diabetic patients to match the positive class size. Based on our hyperparameter tuning, the optimal number of trees for the diabetes model was 80. For final assessment of generalization, we evaluated this fixed model on the unseen test set, reporting accuracy, sensitivity (recall), specificity, precision, and F1-score by using 10-fold cross validation. All steps were implemented in scikit-learn 1.1.3 with Python 3.10.

Results:

The model achieved a test accuracy of 67%, with a sensitivity of 0.624, specificity of 0.680, precision of 0.293, and an F1 score of 0.398 for stroke and test accuracy of 66.2%, with a sensitivity of 0.615, specificity of 0.674, precision of 0.311, and an F1 score of 0.413 for diabetes.

Model comparison and feature importance is shown in Figure 1-4 for stroke and diabetes.

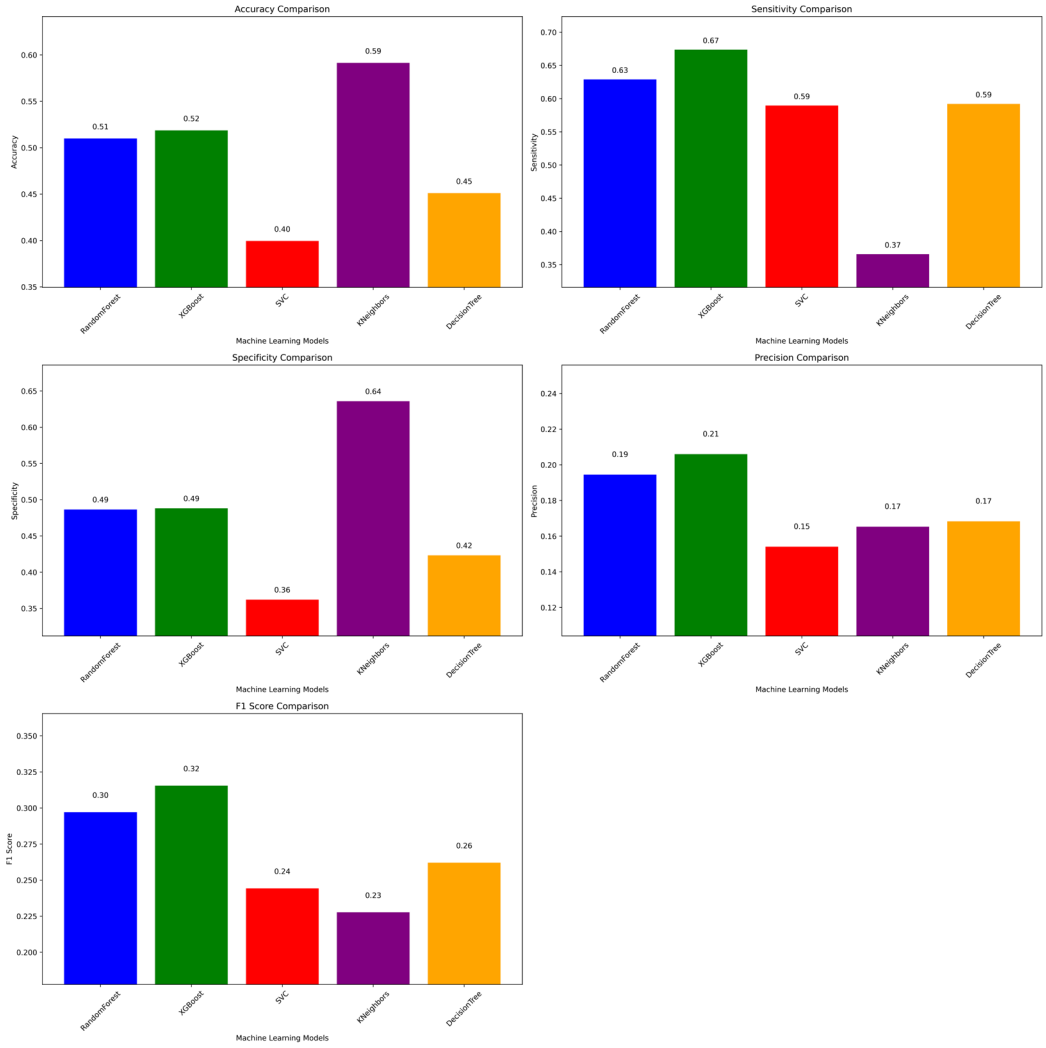

Figure 1. model comparison of stroke prediction

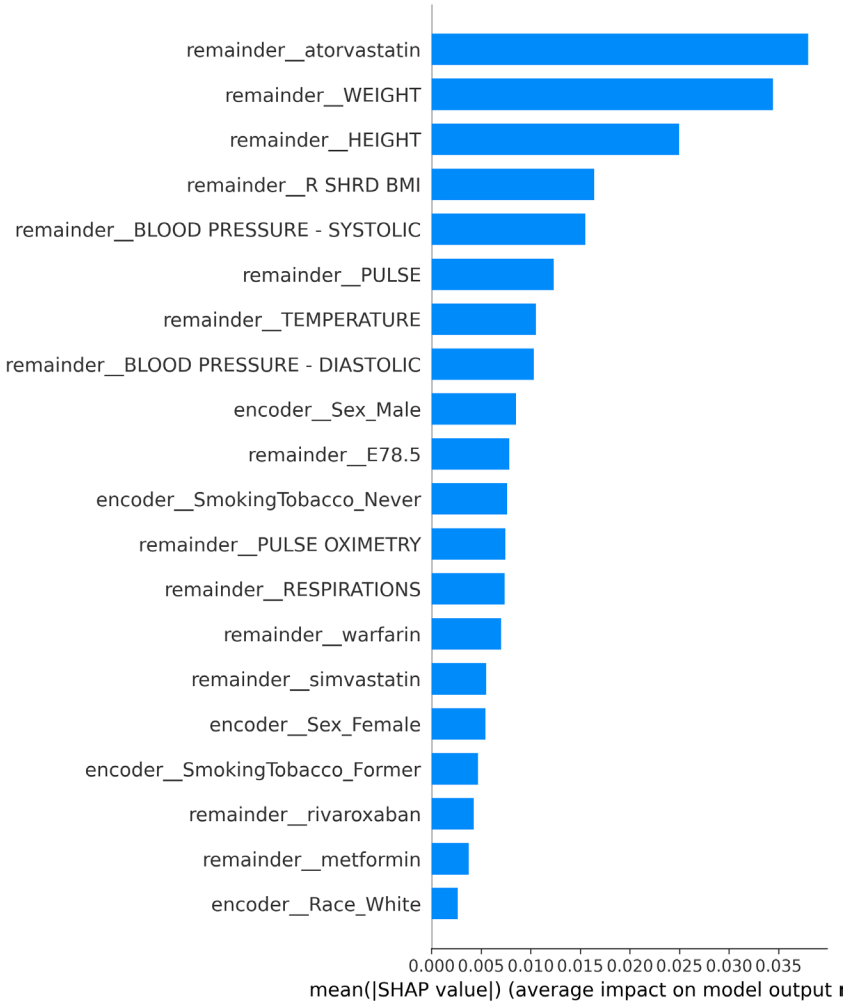

Figure 2. Feature importance of stroke prediction

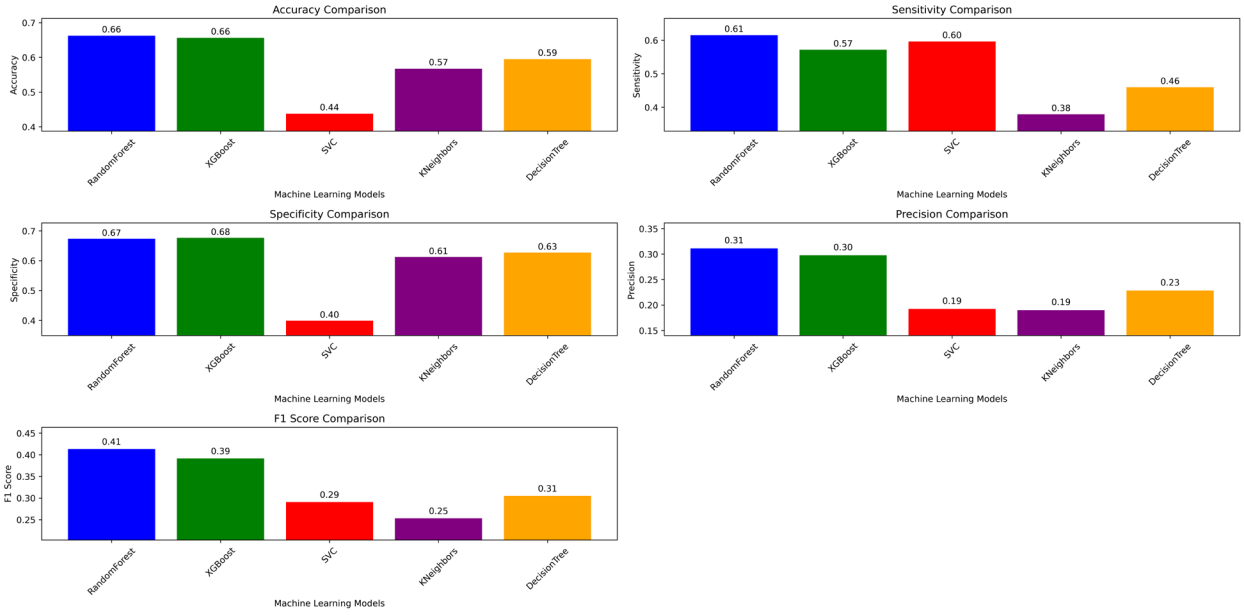

Figure 3. model comparison of diabetes prediction

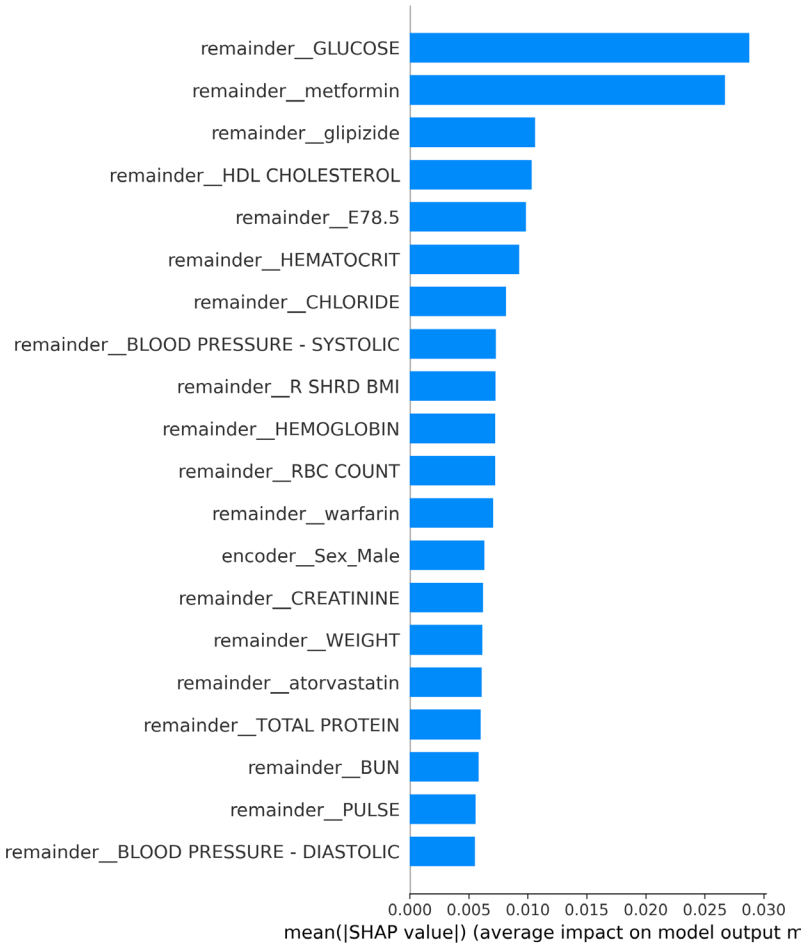

Figure 4. Feature importance of diabetes prediction

Discussion

This project aimed to develop a machine learning model capable of predicting five-year stroke risk in patients newly diagnosed with atrial fibrillation (AF). At the outset, the dataset was constructed using all available patient data, with stroke status serving as the primary reference point for inclusion. This approach allowed for a large initial sample size but lacked methodological rigor, as it did not account for the temporal relationship between AF diagnosis and the clinical data used for prediction. As the project progressed, the methodology was refined to include only data collected around the time of the first AF diagnosis, thereby improving the interpretability and clinical relevance of the dataset. However, the initial attempt to use the AF diagnosis date as a strict cutoff resulted in minimal available data for many patients, as key labs and vitals were often recorded shortly after the diagnosis. To address this, the reference window was expanded to include data from up to one month following the AF diagnosis, which significantly improved data completeness and model input quality. Additionally, we balanced our data for training the ML models. Using the unbalanced data consistently resulted in worse results than when we balanced the data. Despite these

improvements, the model continued to struggle with consistent stroke prediction, underscoring the limitations of the available data. Many patients lacked comprehensive lab or ECG data, and the dataset as a whole was not curated with stroke prediction as a primary objective. This made it difficult to identify strong, consistent predictors of stroke risk. Future work should focus on building a more targeted dataset, ideally with a curated list of approximately ten clinically relevant predictors, rather than relying on over 80 features with varying degrees of completeness. Such a focused approach would likely yield more interpretable and clinically actionable models.