Stroke Prediction Model

1st Danny Li
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
dl953@drexel.edu

2nd Tien Nguyen
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
thn44@drexel.edu

3rd Emily Wang
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
ew552.drexel.edu

***Abstract*—Machine learning techniques are being increasingly adapted for use in the medical field to predict diseases. Stroke is a common and frequently occurring clinical disease, which seriously threatens people's health. Thus, accurate prediction of stroke is highly valuable for early intervention and treatment. Specifically, we consider the common problems of data imputation, data preprocessing, feature selection and prediction in final datasets. We propose an effective feature selection entropy algorithm that will help us choose the top features to compare with related works. Using data collected from International Stroke Trial database, we preprocessed the data and implemented Logistic Regression, Naïve Bayes, and Decision Tree to train and test the data. Furthermore, we used cross validation and random forest models in attempt to improve our accuracy and efficiency statistics. Our work has implemented these models with distinctive features datasets and found random forest produced the best results with the unbalanced dataset with a precision of 0.977 and a recall of 1. As for the balanced dataset, logistic regression provided the best results with 0.699 accuracy, 0.723 precision and 0.714 recall.**

I. Introduction/Related Work

Cardiovascular diseases are a group of disorders of heart and blood vessels, commonly referred to as heart disease and stroke, are the world's number one killer according to the World Heart Federation [1]. Being able to predict stroke will help start early treatment for potential stroke victims. Early treatment will provide greater chance of survivability and recovery. In addition, it can reduce the likelihood of permanent disability and the need of extensive rehabilitation. To help with stroke prediction, various groups of scientist and medical professionals have been collecting and analyzing data on different factors to identify effective predictors of stroke. The International Stroke Trail Database [2] conducted a study to determine if early administration of aspirin, heparin, both or neither influenced the clinical course of acute ischemic stroke. For the research, the International Stroke Trail Database collected features such as age, sex, walking symptoms, atrial fibrillation, face deficit and other physiological features of each patient. Framingham Study [3,4] is another study that reported a list of stroke risk factors including age, diabetes mellitus, cigarette smoking, prior cardiovascular disease, and other predictor features of stroke.

Using collected data and predicting stroke through a machine learning approach is the goal of our research, but it is not a new concept. The use of machine learning to predict and identify stoke has already been explored by multiple groups of researchers. Using the data set from the International Stroke Trial Database [2], "Stroke Prediction Using SVM" by R.S Jeena and Sukesh Kumar [5] implemented support vector machine (SVM) with different kernel functions to predict stroke. SVM is a supervised learning method that finds the optimal hyperplane that distinctly classifies the data points. To deal with non-linearity and higher dimensions, kernels are used to add dimension to the data set [5]. Using different types of kernels will transform the data set in different ways and produce models with different levels of separability. The more separated the data is, the better the results SVM will produce in classification. The research computed the parameters sensitivity, specificity, accuracy, precision and F1 to evaluate the performance of the various functions of SVM classifier. The final results from Jeena and Kumar's research showed linear kernel producing the best results with an accuracy of 91% [5]. The origin of our data set is the same as Jeena and Kumar's research, but we will preprocess the data differently and use different types of classifications algorithm to compare the results.

"Machine Learning Approach to Identify Stroke Within 4.5 Hours" by Hyunna Lee and et al. is another work that uses machine learning to identify stroke. Hyunna Lee and et al.'s research analyzed diffusion-weight imaging (DWI) and fluid attenuated inversion recovery (FLAIR) magnetic resonance imaging to identify patients within the recommended time window for thrombolysis, a type of stroke [6]. Their method was to extract 89 vector features from each imaging sequence and train them with logistic regression, support vector machine and random forest. The results showed 75.8% sensitivity for logistic regression and random forest, 72.7% sensitivity for SVM and 48.5% sensitivity for human reading. Human reading is where two stroke neurologists did visual assessment on the DWI and FLAIR images of the patients. Based on the sensitivity results, the research concluded that the ML approach may be feasible and useful when it comes to identification of patients with stroke within 4.5 hours of symptom onset. However, further research is required to evaluate the applicability the ML algorithms to other patient populations [6].

For our research, we will also be using the dataset collected by the International Stroke Trail Database [2]. Our approach begins with preprocessing the data to select features that are predictors of stroke and removing patients that do not have the proper data. The data set used is unbalanced and mainly contains samples of patients that had a stroke. We will train and validate both the balanced and unbalanced data with binary logistic regression, Naïve Bayes classifier, decision tree classifier using the ID3 algorithm, and random forest. The data preprocessing steps and the classifier algorithms we use will be explained in more detail in the next section. After explaining our method and approach, the paper will show and discuss the results of our work before making a conclusion and explaining what future work can be done to improve the method of using machine learning to predict stroke.

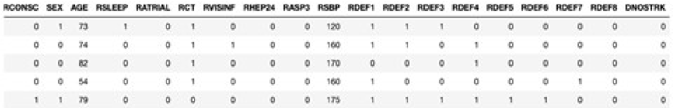
II. Method

A. Overall Approach

For our research, we are using the dataset from the International Stroke Trial Database that contains 122 features and 19435 samples. The first step in our method is to preprocess the dataset to contain features that are predictors of stroke and get rid any samples that do not contain the necessary data. A more detailed explanation of the data preprocessing step is explained in the following section Data Preprocessing section. After preprocessing the dataset, we will use machine learning techniques to train and validate the dataset. We will use binary logistic regression, Naïve Bayes classifier and decision tree (ID3) classifier to train and validate the dataset. . Random forest will also be used on the unbalanced dataset to see if the results can improve. In addition, cross validation will be used on balanced datasets in attempt to improve the results since the balanced dataset contain smaller amount of samples (624 samples). Each of these machine learning techniques are different and details on how they perform are explained in later sections. The Efficiency Analysis section will explain how the efficiency each algorithm are computed and what types of results will be compared.

B. Data Preprocessing

Initially, the dataset from International Stroke Trail Database contains 122 features and 19,435 samples. Not all 122 features are useful features when it comes to predicting stroke. For example, features such hospital id number, discharge dates and other features that are not predictors of stroke are not needed for our research. As result, we delete features that are not predictors of stroke and features that have empty or missing value. After filtrating the data, we end up with the 18 features, shown in Table 1, as our basic dataset which has 13,079 samples.



RCONSC	SEX	AGE	RSLEEP	RATRIAL	RCT	RVISINF	RHEP24	RASP3	RSBP	RDEF1	RDEF2	RDEF3	RDEF4	RDEF5	RDEF6	RDEF7	RDEF8	DNOSTRK
0	1	73	1	0	1	0	0	0	120	1	1	1	0	0	0	0	0	0
0	0	74	0	0	1	1	0	0	160	1	1	0	1	0	0	0	0	0
0	0	82	0	0	1	0	0	0	170	0	0	0	1	0	0	0	0	0
0	0	54	0	0	1	0	0	0	160	1	0	0	0	0	0	1	0	0
1	1	79	0	0	0	0	0	0	175	1	1	1	1	1	1	0	0	0

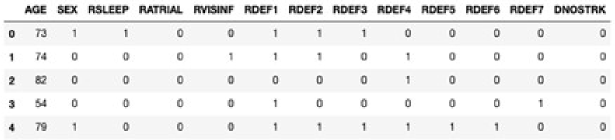
Fig. 1. Subset of 18 Feature Dataset.

.

TABLE I
Meaning of TP, TN, FP and FN

Feature	Feature Meaning
RCONSC	Conscious state of randomization (Fully Alert = 0, Drowsy = 1, Unconscious = 2)
SEX	Male =0, Female = 1
AGE	Age of patient in years
RSLEEP	Symptoms noted on waking (Y=1, N=0)
RATRIAL	Atrial fibrillation (Y=1, N=1)
RCT	CT before randomization (Y=1, N=1)
RVISINF	Infarct visible on CT (Y=1, N=1)
RHEP24	Heparin within 24 hours prior to randomization (Y=1, N=0)
RASP3	Aspirin within 3 days prior to randomization (Y=1, N=0)
RSBP	Systolic blood pressure at randomization(mmHg)
RDEF1	Face deficit (Y=1, N=0)
RDEF2	Arm/hand deficit (Y=1, N=0)
RDEF3	Leg/foot deficit (Y=1, N=0)
RDEF4	Dysphasia (Y=1, N=0)
RDEF5	Hemianopia (Y=1, N=0)
RDEF6	Visuospatial disorder (Y=1, N=0)
RDEF7	Brainstem/cerebellar signs (Y=1, N=0)
RDEF8	Other deficit (Y=1, N=0)
DNOSTRK	Not a stroke (Y=1, N=0)

To compare the results of our research with R.S Jeena and Sukesh Kumar paper's result, we decided to use the same amount of features they did with their research. We decided to take two routes to decide what 12 features should be selected from the 18 features. The first 12 feature dataset will contain the same features as the ones used in "Stroke Prediction Using SVM" by R.S Jeena and Sukesh Kumar [5] and are shown in Figure 2.

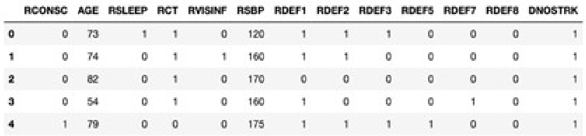


	AGE	SEX	RSLEEP	RATRIAL	RVISINF	RDEF1	RDEF2	RDEF3	RDEF4	RDEF5	RDEF6	RDEF7	DNOSTRK
0	73	1	1	0	0	1	1	1	0	0	0	0	0
1	74	0	0	0	1	1	1	0	1	0	0	0	0
2	82	0	0	0	0	0	0	0	1	0	0	0	0
3	54	0	0	0	0	1	0	0	0	0	0	1	0
4	79	1	0	0	0	1	1	1	1	1	1	0	0

Fig. 2. Subset of 12 Paper Feature Dataset.

.

For the second 12 feature dataset, we calculated the entropy for each of 18 features. Entropy is the measure of disorder or uncertainty. Since the goal of our models is to reduce uncertainty, we selected the 12 features with the lowest entropy for the second data set. The 12 features with the lowest entropy are shown in Figure 3.



	RCONSC	AGE	RSLEEP	RCT	RVISINF	RSBP	RDEF1	RDEF2	RDEF3	RDEF5	RDEF7	RDEF8	DNOSTRK
0	0	73	1	1	0	120	1	1	1	0	0	0	1
1	0	74	0	1	1	160	1	1	0	0	0	0	1
2	0	82	0	1	0	170	0	0	0	0	0	0	1
3	0	54	0	1	0	160	1	0	0	0	1	0	1
4	1	79	0	0	0	175	1	1	1	1	0	0	1

Fig. 3. Subset of 12 Entropy Feature Dataset.

.

After the first preprocessing, we have three datasets:

- 18-features: the basic dataset;
- 12-paper-features: has 12 features that is similar to what the research paper used [5];

- 12-features: has 12 features that has lowest uncertainty based on entropy values;

The 18-features dataset is imbalanced with 97.6% sample are strokes. In fact, with 13,079 samples, 12,767 samples are stroke, and the rest 312 samples are no-stroke. Figure 4 show the imbalance of this dataset.

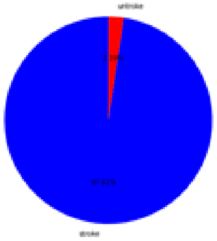


Fig. 4. Dataset Distribution.

.

Since the 12-paper-features and the 12-features were created from the 18-features, these two datasets are also imbalanced with the same number of samples. To create a balanced dataset, 312 stroke samples of the original dataset were randomly selected and combined with 312 no-stroke samples. Two balanced datasets were generated: one for the 12-paper-features dataset and one for the 12-features dataset. Each of them has 12 features and 624 samples. Therefore, we have a total of five dataset for the entire project.

- 12-paper-features-balanced: a balanced dataset from 12-paper-features dataset.
- 12-feature-balanced: a balanced dataset from 12-features dataset.

C. Binary Logistic Regression

The first classification technique we will be using on the data to predict stroke is binary logistic regression. Binary logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories [7]. The independent variables in this research will be predictor of stroke features and the dependent variable categories are stroke or no stroke. Logistic regression works by forming a best fitting equation or function using the maximum likelihood method to maximize the probability of classifying the observed data into the appropriate class given the regression coefficients [7]. The log likelihood can be represented by the equation below:

$$J = y \ln(\hat{y}) + (1 - y) \ln(1 - \hat{y}) \tag{1}$$

In the equation, J represent the likelihood, y represents the target values and $\hat{y}$ represent the predicted/computed values. To maximize the likelihood, we first take the derivative of the log likelihood equation.

$$\frac{\partial J}{\partial \mathbf{w}} = (y - \hat{y})\mathbf{x}^T \tag{2}$$

The derivative is taken with respect to $\mathbf{w}$, which represents the vector of all regression coefficients and the bias (b). After determining the derivative, we use the gradient ascent rule to iteratively update the regression parameter coefficient values as we train the data by following the following equation.

$$\mathbf{w} := \mathbf{w} + \eta \left(\frac{\partial J}{\partial \mathbf{w}} \right) \tag{3}$$

For our implementation of logistic regression, the termination criteria are based on the change in mean log of likelihood and the number of epochs ran during iteration.

D. Naive Bayes Classifier

The next type of classifier we use for stroke prediction is Naïve Bayes classifier. The Naïve Bayes classifier is a probabilistic machine learning model based on the Bayes theorem [8].

$$P(y|x) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})} \tag{4}$$

The Bayes theorem finds the probability of y happening when x occurs. In the equation, y represents the final class, which is stroke or no stroke, and $\mathbf{x}$ represents all the features or predictors of stroke. $P(y)$ is referred to as the prior and represents the probability of a certain class and $P(x|y)$ represents the probability of class y generating the observed features x.

For the Naïve Bayes approach, we assume that the features are conditionally independent and rewrite $P(x|y)$ as [9]:

$$P(\mathbf{x}|y) \approx \prod_{j=1}^{D} P(x_j|y) \tag{5}$$

Substituting the $P(x|y)$ into the previous equation we would end up with:

$$P(y|x) = \frac{P(y) \prod_{j=1}^{D} P(x_j|y)}{P(\mathbf{x})} \tag{6}$$

For our method, we are using norm probability density function or the Gaussian Naïve Bayes to calculate $P(x_j|y)$. This method assumes the values of the dataset sample for a Gaussian distribution or normal distribution. As a results $P(x_j|y)$ can be determined by:

$$P(x_j|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\ \sigma_y^2} \right) \tag{7}$$

It is important to note that the denominator of the $P(y|x)$ equation is the same for all entries in the dataset and can be removed to form the following proportionality [8].

$$P(y|\mathbf{x}) \propto P(y) \prod_{j=1}^{D} P(x_j|y) \tag{8}$$

Finally, to determine the classification of a patient or sample we would find the class with the maximum probability for it.

$$y = argmax_y P(y) \prod_{j=1}^{D} P(x_j|y) \tag{9}$$

E. Decision Tree (ID3) Classifier

The third classification we are using is the decision tree classifier using the ID3 algorithm. A decision tree classifier is a predictive modeling approach using a tree model to classify a sample data. Tree models are made up of nodes that tests an attribute, leaves that represent class labels, branches that represent attribute values that lead to nodes or class labels. An example of a decision tree is shown in Figure 1.

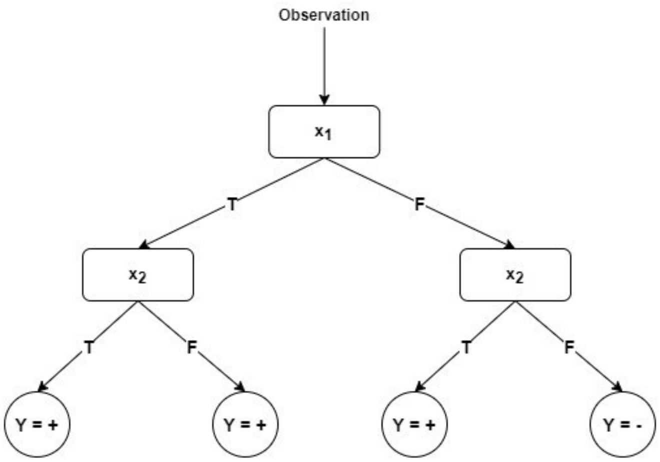
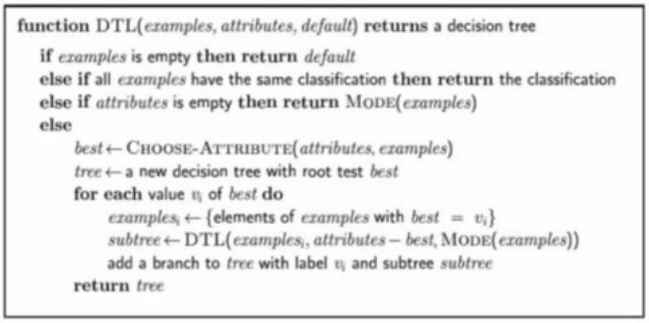


Fig. 5. Decision Tree Example.

In Figure 1, an observation is inputted into a decision tree and the first node represents feature x_1. Feature x_1 will be tested and the result of it will determine which branch path to take. Then feature x_2 will be tested to determine what the final class label is for the observation.

The decision tree for our research is created based on the greedy algorithm. Figure 2 shown below is a pseudo-code that represents the greedy algorithm shown in equation 10[10].



```
function DTL(examples, attributes, default) returns a decision tree
    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value $v_i$ of best do
            $examples_i$ ← {elements of examples with best = $v_i$}
            subtree ← DTL($examples_i$, attributes − best, MODE(examples))
            add a branch to tree with label $v_i$ and subtree subtree
        return tree
```

Fig. 6. Greedy Algorithm Pseudo-code.

Within the greedy algorithm, we decide the best attributes at each level of the tree based on the entropy. Entropy is the measure of randomness of a system. For a system with n possible states, v_1,. . . ,v_n and their probabilities of occurrence, $P(v_1)$,. . . ,$P(v_n)$, we can compute the entropy as [10].

$$H(P(v_1), ..., P(v_n)) = \sum_{i=1}^{n} (-P(v_i) \log_n P(v_i)) \tag{10}$$

After creating the decision tree based on our training data, each validation data will run through the decision tree and be classified based on which leaf node they end up in.

F. Cross Validation

One of the methods to overcome overfitting in machine learning model is using cross-validation to train on more data. In real life, generating more data may not be realistic. Cross-validation can be used to have the model be trained on more data using the same dataset. There are a few types of cross-validation. The one used in this research called k-Folds. The dataset will be divided up into k-parts. The model will be trained using $k-1$ parts and the remaining part will be used for validation. With cross-validation, multiples systems are built along with their validation statistics. The final statistics of the model will be the average of the system statistics. The equation below explains the how the accuracy is calculated for a k-folds model.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} accuracy_i \tag{11}$$

G. Efficiency Analysis

The efficiency of each classifier will be analyzed based on the error rate. For each classifier, the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are obtained to help calculate metrics such as precision, recall, f-measure, and accuracy to evaluate and compare the classifiers.

TABLE II
MEANING OF TP, TN, FP AND FN

True Positive (TP)	Classified as positive and is a stroke patient
True Negative (TN)	Classified as negative and is not a stroke patient
False Positive (FP)	Classified as positive and is not a stroke patient
False Negative (FN)	Classified as negative and is a stroke patient

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{15}$$

Precision measures the percentage of classified positives or stroke cases that were actually positive. Recall measures the percentage of correctly identified true positives or stroke cases. F-measure is a weighted harmonic mean of precision and recall. Accuracy is the amount of correctly identified classifications over the total number of classifications. These four types of measurements are calculated to each classifier and compared to determined which classifier performs the best with the given dataset.

H. Random Forest

Random forest is an ensemble approach for classification. This classifier builds multiple trees by selecting a random set of features. The output for random forest is the class selected by the most tree. In random forest classifier, there are three main parameters needed to be specified: the size of samples M to be selected without replacement, the number of trees in the forest N, and the number of features to be selected randomly from remaining p variables. While M and N could be selected based on our preferences, m is defined as $m = \sqrt{p}$ for our project since it returned the desired statistics. Random Forest has many advantages such as:

- It can produce data of very high dimensions (many features) without dimensionality reduction and feature selection.
- It can determine how important feature is.
- It is easy to implement and can avoid overfitting.
- For unbalanced dataset, it can balance out errors.
- If a significant portion of the feature is missing, accuracy can still be maintained.

III. Results/Evaluation

A. Train and Evaluate on Imbalance Datasets

The developed Logistic Regression, Naïve Bayes, and Decision Tree (ID3) classifiers were firstly trained and evaluated using three datasets: 18-features, 12-paper-features, and 12-features. The statistics of validation for of each model is reported in Table III.

For Logistic Regression classifier, precision values were equally 0.975 for the 18-features, 12-paper-features, and 12-features datasets, respectively. The recall values are approximate 1 for all three datasets. Based on these two values, it can be concluded that the Logistic Regression is nearly 100% successful classify the datasets. In fact, A value of 0.975 means that 97.5% of the predicted stroke is correct. A value of 1 for recall means that there were no FN in the result. With this high precision and recall, a similar value for accuracy is expected. However, it is interesting that the statistics values for all three datasets are equal to one another although these datasets have different features. While the statistics values are suspicious, more investigation in the implementation is needed.

For Naïve Bayes classifier, precision values for all three datasets were approximately 0.979 which means that 97.9% of the patients who were predicted having stroke is correct. The recall values ranged from 0.940 to 0.960 means that only 94% to 96% of people who have stroke are predicted correctly. Additionally, the developed Naiver Bayes classifier could predict from 92.3% to 94.1% accuracy depending on the dataset. Naive Bayes classifier produced the lowest statistics comparing to Logistic Regression and Decision Tree classifiers. For Naïve Bayes, we assume that the features are conditionally independent. However, this may not be true. For example, the "age" feature and the "atrial fibrillation" are not independent. Atrial fibrillation increases with age especially for individuals between the age of 65 and 85 years old [13]. This inaccurate assumption may lead to a lower performance of Naïve Bayes.

For Decision Tree classifier, precision values for all three datasets are 0.976 which is slightly lower than the results from Naïve Bayes classifier. However, the recall, f-measure, and accuracy values were higher comparing to these values from Naïve Bayes classifier. For example, the recall value for the 12-paper-features here is 99.9% which means that 99.9% people who have stroke are correctly predicted while it was only 96% for the Naïve Bayes classifier. f-measure and accuracy values are also reported to be 0.987 and 0.973, respectively.

Additionally, the developed Decision Tree classifier produces the same root, infarct visible on computed tomography (CT) for all three datasets. Based on the developed trees, "infarct visible on CT" is the first feature to predict if a person is at risk of having a stroke. Infarction is an area of tissue that is dead or dying due to loss of blood supply. According to a research by Wardlaw and colleagues [11], the visible infarction on CT is could be an adverse prognostic indicator for stroke which confirmed the root of the developed decision trees is reasonable. If the return value from the root is "Y" (yes), the "leg/foot deficit" would be the next feature to be considered. If the return value from the root is "N" (no), the "arm/hand deficit" would be the next feature to be considered. In fact, according to CDC, arm and leg deficit are one of the top five signs to identify stroke in men and women [12].

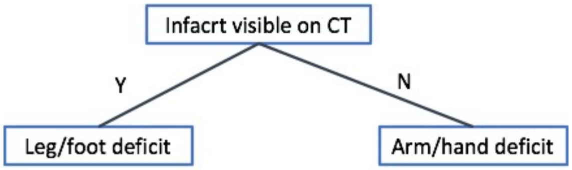


Fig. 7. Root of Decision Tree for all Three Datasets.

From these statistics values, Logistic Regression classifier is concluded to produce the highest precision, recall, f-measure, and accuracy for all three datasets. Also, these classifiers perform best on the 12-paper-features dataset.

B. Ensemble Technique: Random Forest

Normally, achieving 97.5% classification accuracy would be cause to celebration. However, the class distribution is imbalanced with 97.6% being class 1. Therefore, 97.6% is actually the lowest acceptable accuracy for these imbalanced datasets. Random Forest was introduced with the hope to improve the accuracy. Random Forest classifier was trained and evaluated on all previous three datasets. Since they gave similar statistics results at the end, only the statistics for the

TABLE III
VALIDATION STATISTICS RESULTS FOR THE THREE DATASETS

Model	Dataset	Precision	Recall	f-measure	Accuracy
Logistic Regression	18 features	0.975	1	0.987	0.975
	12 paper features	0.975	1	0.987	0.975
	12 features	0.975	1	0.987	0.975
Naive Bayes	18 features	0.980	0.940	0.960	0.923
	12 paper features	0.979	0.96	0.969	0.941
	12 features	0.979	0.946	0.962	0.928
Decision Tree (ID3)	18 features	0.977	0.985	0.981	0.963
	12 paper features	0.975	0.999	0.987	0.973
	12 features	0.976	0.997	0.986	0.973

12-paper-features is reported in Table IV. As stated earlier, 2/3 of 13,079 samples dataset was used for training. The sample size of the Random Forest is required to be smaller than 8,719 samples. Samples size of 6,000 and 8,000 were implemented on the classifier. The number of decision trees was varied from 1 to 10 for the sample size of 6,000, and from 1 to 500 to the sample size of 8,000.

TABLE IV
STATISTICS RESULTS FOR RANDOM FOREST CLASSIFIER

Sample	Number	Precision	Recall	f-measure	Accuracy
Size	of Trees				
6000	1	0.977	0.999	0.988	0.976
	10	0.977	1	0.988	0.977
8000	1	0.977	0.999	9.988	9.975
	10	0.977	1	0.988	0.977
	100	0.977	1	0.988	0.977
	200	0.977	1	0.988	0977
	300	0.977	1	0.988	0.977
	500	0.977	1	0.988	0.977

For both sample size, the statistics values are similar with 10 trees in the random forest. The statistics results were constant with an increasing in number of trees. The highest and constant values of precision, recall, f-measure, and accuracy are 0.977, 1, 0.988, and 0.977, respectively. A precision of 0.977 means that 97.7% of the patients who were classified having stroke is correct. A recall value of 1 means that 100% of the patients who have stroke were classified correctly. F-measure of 0.988 is the harmonic mean of the precision and recall. In the case of imbalance of binary class datasets, f-measure is really useful since it accounts for the costs of false positive and false negative. An accuracy of 0.977 means that 97.7% samples were classified correctly. This number is slightly greater than the lowest acceptable accuracy, and random forest classifier is the only classifier achieved this target.

C. Train and Evaluate on Balanced Datasets

The developed Logistic Regression, Naïve Bayes, and Decision Tree classifiers are also trained and evaluated with the balanced datasets. Each of the balanced datasets has 12 figures and 624 samples with a ratio of 1:1 stroke and no-stroke. The statistics results for training and validation of each classifier on the 12-paper-features-balanced is reported in Table 3.

TABLE V
TRAINING AND VALIDATION RESULTS FOR THE BALANCED DATASET

Model	Dataset	Precision	Recall	f-measure	Accuracy
Size	of Trees				
Logistic Regression	Training	0.665	0.662	0.664	0.673
	Validation	0.723	0.714	0.719	0.699
Naive Bayes	Training	0.722	0.528	0.610	0.680
	Validation	0.730	0.470	0.571	0.611
Decision Tree (ID3)	Training	0.716	0.906	0.800	0.815
	Validation	0.504	0.652	0.569	0.562

For Logistic Regression classifier, the statistics results for validation in general is higher than the statistics results for training. For example, the precision values for validation and training are 0.723 and 0.665, respectively. The accuracy for validation and training are 0.699 and 0.673, respectively. However, an opposite result was observed for the Naïve Bayes and Decision Tree classifiers. The statistics results for the training datasets are higher than the respective validation datasets for both classifiers which indicates overfitting error on these two classifiers. For example, Naive Bayes classifier returned accuracy values of 0.611 and 0.680 for validation and training, respectively. The overfitting error is greater on Decision Tree classifier. The differences between the training and validation statistics are even higher. For example, here, the accuracy values are 0.562 and 0.815 for validation and training, respectively.

As mentioned in the introduction, we used the same database from the International Stroke Trial and the same 12 features as “Stroke Prediction Using SVM” by R.S Jeena and Sukesh Kumar [5]. Even though we used the same database and 12 features, we had different number of samples after preprocessing and used different machine algorithm to train and validate the data. Jeena and Kumar’s paper demonstrated SVM with linear kernel producing the best results with 91% accuracy. Our research produced better results using random forest on imbalanced dataset and logistic regression on balanced dataset with 97.7% and 97.5% accuracy, respectively. Though we produced higher accuracy levels, it is important to note that our dataset was different after preprocessing and Jeena’s research does not mention if their data after preprocessing is balanced or imbalanced.

Overall, due to smaller size of the dataset, the statistics results for all three classifiers on the balanced dataset are lower than the statistics on the imbalance one. Overfitting error was reported for Naïve Bayes and Decision Tree.

To overcome overfitting error on the balanced dataset, cross-validation was introduced. The number of k-folds was varied from 5 to 500. The validation accuracy values for all three classifiers are reported in Table VI.

TABLE VI
Training and Validation Results for the Balanced Dataset

K-fold	Logistic Regression	Naive Bayes	Decision Tree
5	0.513	0.569	0.516
10	0.590	0.607	0.555
50	0.622	0.657	0.578
100	0.669	0.665	0.585
500	0.682	0.674	0.594

According to Table VI, the accuracy of all classifiers increased while number of k-folds increased. While k-fold increased from 5 to 500, the validation accuracy increased from 0.513 to 0.590, from 0.569 to 0.674, and from 0.516 to 0.594 for Logistic Regression, Naïve Bayes, and Decision Tree classifier. Here, the highest validation accuracy, 0.682, for the balanced dataset was achieved with Logistic Regression classifier.

The accuracy for Naïve Bayes and Decision Tree classifiers were improved from 0.611 to 0.674 and from 0.562 to 0.594, respectively (Table V). However, the accuracy from cross-validation is still lower than the validation accuracy for the Logistic Regression classifier without cross-validation (Table V). Therefore, cross-validation only improved the accuracy for Naïve Bayes and Decision Tree classifier.

IV. Conclusion/Analysis

Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest classifiers were trained and evaluated on the imbalance dataset. Random Forest classifier returned the highest statistics with a precision of 0.977 and a recall of 1. Logistic Regression, Naïve Bayes, and Decision Tree were trained and evaluated on the balanced dataset. The statistics results were much lower comparing to the imbalanced dataset due to a smaller in dataset size. Logistic Regression returned the highest statistics results for the balanced dataset. Cross-validation was also implemented on the balanced dataset, and it improved the accuracy while minimizing overfitting for Naïve Bayes and Decision Tree classifiers.

As mentioned in the introduction, we used the same database from the International Stroke Trial and the same 12 features as “Stroke Prediction Using SVM” by R.S Jeena and Sukesh Kumar [5]. Even though we used the same database and 12 features, we had different number of samples after preprocessing and used different machine algorithm to train and validate the data. Jeena and Kumar’s paper demonstrated SVM with linear kernel producing the best results with 91% accuracy. Our research produced better results using random forest on imbalanced dataset with 97.7% accuracy. Though we produced higher accuracy levels, it is important to note that our dataset was different after preprocessing and Jeena’s research does not mention if their dataset after preprocessing is balanced or imbalanced.

In conclusion, Random Forest classifier produced the highest reasonable statistics results for the project. In fact, this classifier generated zero FN. It is the only classifier that achieve an accuracy that is higher than the minimum acceptable accuracy for an imbalance dataset. Therefore, Random Forest is so far the best model to predict if a person is at risk of stroke.

V. Future Work

We have explored three different types of classification techniques on the data set obtained from the International Stroke Trial database. For future work, we can explore other types of machine learning approaches on the same dataset to determine if there are better methods to predict stroke. Examples of other machine learning classification techniques that can be used, include artificial neural networks and k-nearest neighbors. It is also feasible to explore other data balancing techniques to see which method would produce the best results with the unbalanced dataset from the International Stroke Trail database.

Other than exploring other machine learning techniques, different datasets can also be analyzed and trained for comparison. Different datasets can contain different predictors of stroke, which can produce different results since different predictors can have different levels of correlation with stroke. In addition, analyzing different or additional groups of patient data can help with overfitting and produce better results.

References

[1] “What Is CVD?” World Heart Federation, 6 May 2021, world-heart-federation.org/what-is-cvd/.

[2] Sandercock, P.A., Niewada, M., Członkowska, A. et al. The International Stroke Trial database. Trials 12, 101 (2011). https://doi.org/10.1186/1745-6215-12-101

[3] T. R. Dawber, G. F. Meadors, and F. E. Moore. Epidemiological approaches to heart disease: The Framingham study. American Journal of Public Health and the Nation’s Health, 41:279–286, March 1951.

[4] P. A. Wolf, R. B. D’Agostino, A. J. Belanger, and W. B. Kannel. Probability of stroke: a risk profile from the Framingham study. Stroke, 22:312–318, March 1991

[5] R. S. Jeena and S. Kumar, ”Stroke prediction using SVM,” 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 600-602, doi: 10.1109/ICCICCT.2016.7988020.

[6] Lee, Hyunna, et al. “Machine Learning Approach to Identify Stroke Within 4.5 Hours.” Stroke, vol. 51, no. 3, 28 Jan. 2020, pp. 860–866., doi:10.1161/strokeaha.119.027611.

[7] Hua, Cheng, et al. “Binary Logistic Regression.” Companion to BER 642: Advanced Regression Methods, 2021, bookdown.org/chua/ber642_advanced_regression/#course-description.

[8] Gandhi, Rohith. “Naive Bayes Classifier.” Medium, Towards Data Science, 17 May 2018, towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[9] Burlick, Matthew Drexel CS613 Machine Learning Bayesian Classification Power Point Slides

[10] Burlick, Matthew Drexel CS613 Machine Learning Decision Trees Power Point Slides

[11] Wardlaw, J.M. et al. ”Is Visible Infarction On Computed Tomography Associated With An Adverse Prognosis In Acute Ischemic Stroke?”. Stroke, vol 29, no. 7, 1998, pp. 1315-1319. Ovid Technologies (Wolters Kluwer Health), doi:10.1161/01.str.29.7.1315. Accessed 7 June 2021.

[12] ”Stroke Signs And Symptoms — Cdc.Gov”. Cdc.Gov, 2021, https://www.cdc.gov/stroke/signs_symptoms.htm.

[13] Karamichalakis, Nikolaos et al. “Managing atrial fibrillation in the very elderly patient: challenges and solutions.” Vascular health and risk management vol. 11 555-62. 27 Oct. 2015, doi:10.2147/VHRM.S83664