D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220075

Efficient Prediction of Stroke Patients Using Random Forest Algorithm in Comparison to Support Vector Machine

Ritaban Mitra[a] and T. Rajendran[b,1]
[a]*Research Scholar, Dept. Of CSE, Saveetha School of Engineering,*
[b]*Asso. Prof., Dept. Of CSE, Saveetha School of Engineering,*
[a,b]*SIMATS, Chennai.Tamil Nadu, India*

Abstract. The work aims to make an efficient prediction of stroke in patients using several Machine learning modeling techniques and evaluating their performance. The two groups used in this paper are the Random Forest Algorithm (RFA) and the Support Vector Machine(SVM) Algorithm. The dataset implemented and tested consists of over 5000 records of patients' medical and personal records. They were using N = 20 iterations for each algorithm. The G-Power test used is about 80%. The results of our work have given us the mean accuracy of 94.61 on Random Forest and 93.91 on Support Vector Machine Algorithms. The statistically significant difference was obtained by generating independent sample t-tests at 0.015. This work is intended to implement innovative approaches to increase the efficiency of stroke prediction algorithms and improve the accuracy of existing algorithms. The results show that the Random Forest Model performs higher than Support Vector Machines.

Keywords. Innovative Stroke Prediction, Machine learning, Data Science, Random Forest Algorithm, Support Vector Machine Algorithm, Statistical Analysis.

1. Introduction

Stroke is the second biggest reason of mortality globally, as per the WHO report, accounting for 11% of fatalities yearly. A stroke is a medical emergency that causes damage to the brain due to a shortage of blood supply, causing brain cells to die. This research paper will explore stroke conditions and use a Machine learning approach to solve this problem and develop an Innovative Stroke Prediction technique in patients [1]. Over the years, as computers have become more powerful, their ability to support research work in the medical domain has also increased. This is a massive benefit to the world as it can combine the power of human intelligence with the potential of computers and gain insights into patterns from Statistical Analysis [2]. This analysis is done using a Data Science driven approach [3]. Applications of the research include clinical prognosis and drug development [2]. The Prediction and classification of heart failure have been made using a Machine learning approach [4]. In paper [5], a

[1] T. Rajendran· Department of Computer Science and Engineering Saveetha School of Engineering Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India, Email:rajendrant.sse@saveetha.com.

combined Machine learning technique is used to identify cerebral stroke for medical assessment based on minimal physiological data and class imbalance. Data Science algorithms such as random forest and the accuracy of Auto-HPO, which is based on deep neural networks, was 71.6 percent. In another study, an explainable model approach was conducted with Random Forest and Support Vector Machines having an accuracy of 78% and 74%, respectively [6]. In the study [7], A maximum accuracy of 95% was attained using machine learning methods such as artificial neural networks, support vector machines, bagging, boosting, and random forests. In the study [8], Support vector machines, decision trees, and random forests were used to model the stroke classification dataset, and the Random Forest Algorithm achieved the best accuracy of 90%. The paper [8] is considered to have the most relevant and accurate results for future Data Science researchers interested in stroke prediction as a data mining approach has been taken to analyze the dataset.

Now this work focuses on this topic. The previously utilized approaches have lower accuracy, are less trustworthy, and are inefficient in terms of stroke prediction. They combine their knowledge and experience with handling various machine learning algorithms to develop innovative solutions for the given problem. The primary goal of this research is to develop a more efficient stroke categorization system by Innovative Stroke Prediction techniques using Machine learning algorithms like RFA and SVM and compare their performance.

2. Materials and Methods

The research has been conducted in the CISCO Lab at SSE, SIMATS. Two Machine learning algorithms based on supervised learning were considered as two separate groups, Random Forests and Support Vector Machines. Two iterations have been performed on each group. Identifying multiple scales by conducting N=20 cycles on every method with a sample size of N = 20 [5]. The G-power test, which enables us to estimate the statistical power of statistical tests, is 80%. A Type-I error, also known as the Alpha error rate, is 0.05, representing the difference between the two methods under consideration. The Research Enrollment ratio is about one based on the input data.

2.1 Dataset Description

The dataset we used for this work is named 'Stroke prediction dataset.' It was made available and accessible on the website Kaggle by user 'fedesoriano,' a data scientist and machine learning practitioner. 11 clinical features in the dataset can be utilized to predict stroke episodes. There are 5110 occurrences of patient information in the dataset.

2.2 Support Vector Machine Algorithm

A Supervised Learning technique called theSupport Vector Machines model is used for performing classification and regression tasks.

Pseudocode for Support Vector Machine Model

```
Input:  TheCollected Dataset
Output: Accuracy Prediction value
    1. Dataset is loaded as input
    2. Dataset is preprocessed and split to train and test
    3. Support Vector Machine classifier is fit to the training dataset
            classifier.fit(xtrain, ytrain)
    4. Test set result is predicted
            Ypred = classifier.predict(Xtest)
    5. Steps 1-3 are repeated for all samples  'l'
    6. Support vector decision boundary is built.
    7. Compute predictscores with various features.
            prediction_score=svm_model.predict(set_parameters, "")
    8. Score-up for every prediction v is calculated.
    9. Obtain final prediction score
```

2.3 Random Forest Algorithm

A Supervised Learning technique called Random Forest can be used for regression and classification. By training a variety of Decision Tree Classification models on diverse sub-samples of the dataset, Random Forests use averaging to improve projected accuracy and control over-fitting. If bootstrap = True (default), the sub-sample size is controlled with the max_samples parameter if bootstrap = True (default). Otherwise, the whole dataset is used to build each tree. The Random Forest Algorithm is an enhanced variant of the Decision Tree.

```
Input:  The Collected Dataset
Output: Accuracy Prediction value
    1. Dataset is loaded as input
    2. Randomly, choose 'x' examples from 'b' data.
    3. Compute the node 'n' from the 'x' data that use the joint distribution.
    4. Nodes are split into child nodes
    5. Steps 1 to 3 are repeated until 'l' number of samples are reached
    6. The random forest has been built.
    7. Compute predict scores with various features.
            prediction_score=rfa_model.predict(set_parameters, "")
    8. Score-up for every prediction v is calculated.
    9. Obtain final prediction score
```

JupyterLab was the python IDE utilized to test the Machine Learning Algorithm. The tool utilized was JupyterLabs with the Python programming language, and the operating system was Windows 10. The testing technique was to divide the data into train and test sets, then use a machine learning classifier to develop and train a model on our data. After training, the predictions are made, and the model's performance is evaluated using the available metrics.

The stroke prediction dataset was gathered via Kaggle. Statistical Analysis techniques were used to prepare data to get some context about the data. Data cleaning techniques are used, such as deleting extraneous attributes and filling in missing values. We can gain some context and valuable insight into the dataset by using data exploration. The Random Forest and Support Vector Machines are compared.

2.4 Statistical Analysis

Independent Sample T-test in the SPSS tool for Statistical Analysis of the Machine learning models was used to evaluate the quality of the study. The independent variables were gender, married, work title, and the resident type, and the dependent variables were heart_disease, BMI, hypertension, average_glucose_levels. The comparison of the Support Vector Machine Algorithm and the Random Forest Algorithm is complete, and the accuracy is found. Hence the Statistical Analysis has been performed, and the observations are noted.

3. Results

The Random Forest Algorithm gave us an accuracy of 95%, and the SVM gave 94% accuracy compared with their accuracy rate. Each method has been iterated 20 times, and the accuracy varies depending on the test size. Due to random changes in the test sizes, a variance in the accuracy is observed, as given in table 1.

Table 1. Accuracy Table of Random Forest and Support Vector Machines

Test Size	0.33	0.55	0.77	0.99
Support Vector Machine Algorithm	94.06	94.35	94.37	94.27
Random Forest Algorithm	95.02	94.98	95.04	95.09

The observed statistical values for these two groups based on critical metrics such as mean accurateness and variance for the Random Forest are 94.96 and 0.29134. The Support Vector Machine Algorithm’s mean accuracy is 94.93, and the standard deviation is 0.45120. The Random forest also obtained a standard error mean rate of 0.6515, whereas the Support Vector Machine Algorithm obtained an error mean rate of 0.10089. The significance value of 0.015 shows that our hypothesis is valid, as given in Table 2.

Table 2. Group Statistics of the mean accuracy and standard deviation.

	RFA, SVM	N	Mean	Standard Deviation	Standard Mean Error
Accuracy	RFA	20	94.6140	.29134	0.6515
	SVM	20	93.9110	.45120	0.10089

Following this, an independent sample test of 10 samples was performed. Random forest obtained a mean difference of 0.70 and a standard error difference of 0.12009. Compared to other algorithms' performance, the Random Forest Algorithm's performance was better than the Support Vector Machine's, as given in Table 3.

Table 3. Independent Samples Test, which is a comparison of accuracy.

		Levene's Test for Equality of Variances (1)	Levene's Test for Equality of Variances (2)	T-test for Equality of Means (3)	T-test for Equality of Means (4)	T-test for Equality of Means (5)
		F	**Sig.**	**Std.Error Difference**	**95% Confidence lower**	**95% Confidence upper**
Accuracy	**Equal variances assumed**	6.529	.015	.12009	.45988	.94612
	Equal variances not assumed			.12009	.45852	.94748

The Innovative Stroke Prediction framework is another name for it. The process for creating a stroke prediction is outlined in the architecture. Data Collection, Preprocess, Exploratory Data Analysis, Modeling Classifier, Deployment, and Evaluation are the steps in the sequence shown in Figure 1.

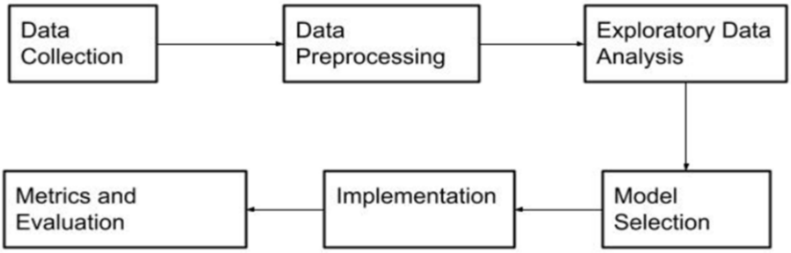

Figure. 1. Machine learning classifier architecture.

The GGraph depicts a bar chart of simple bar mean accuracy, with the Random Forest Algorithm reaching 95 % and the Support Vector Machine Algorithm achieving a 94 %. The 95% error bars represent the variation in the corresponding coordinates of the point. Using The performance of the two algorithms was evaluated using independent - samples t, and a statistical significance is P=0.015 was observed. Further comparing the two methods, the Random Forest Algorithm outperformed Support Vector Machines, as seen in Figure 2.

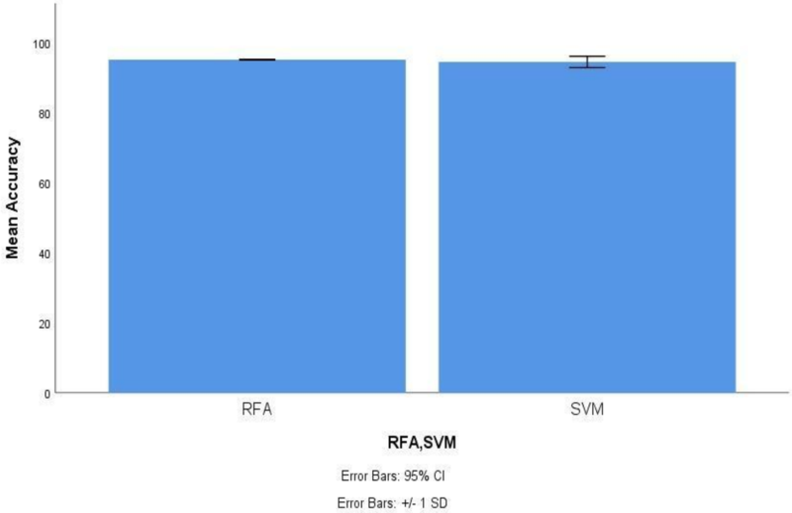

Figure 2. Simple Bar Chart depicts the Mean of Accuracy by RFA and SVM.

4. Discussion

It has been observed that the Random Forest Algorithm performs better in terms of accuracy than Support Vector Machines. The data was gathered through a series of iterations to determine distinct ranges of accuracy rates. Independent samples t-tests are performed on the dataset. In this study of stroke prediction, the Random Forest Algorithm has an accuracy of approximately 95 %, which is higher than that of the Support Vector Machine, which is 94%. Random Forests have a better significance of 0.015 while using the independent samples T-test. The accuracy and SD for the Random Forest Algorithm are observed to be 94.96 and 0.29134 [5], using a missing value imputation and a deep learning model to get an accuracy of 71%. In paper [6], the Support Vector Machine Algorithm obtained an accuracy of 76.5%, and Random Forest achieved an accuracy of 75.2%. Based on the literature review, the Random Forest outperforms Support Vector Machines. By running independent sample tests in IBM's SPSS statistical program, it can be seen that the difference between the two algorithms is statistically significant at p<0.05. The SPSS statistical program also calculates the mean and standard deviation. The paper [1] Support Vector Machine outperformed other algorithms' classification accuracy by 87.8%.

5. Conclusion

In this research work, the results indicate that the proposed Random Forest Algorithm can be used to classify stroke with improved accuracy of 95%. In order to achieve better accuracy, more data would be required. The accuracy of stroke prediction is higher using a Random Forest Algorithm, which is true in the previously conducted studies. One of the major hindrances is that the attributes in the dataset contain fewer data to predict accuracy (%) for stroke classification. The accuracy can be improved by adding additional dependent and independent variables. The dataset provides many parameters for future enhancements that the algorithm may use to increase the accuracy rate. When these features are used, accuracy and precise precision values can be increased.

References

[1] Singh KK, Elhoseny M, Singh A, et al. *Machine Learning and the Internet of Medical Things in Healthcare*. Academic Press, 2021.

[2] Petretta M. Applications of Machine Learning in Medicine. *Biomedical Journal of Scientific & Technical Research*; 20. Epub ahead of print 2019. DOI: 10.26717/bjstr.2019.20.003503.

[3] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019; 380: 1347–1358.

[4] Olsen CR, Mentz RJ, Anstrom KJ, et al. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *Am Heart J* 2020; 229: 1–17.

[5] Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif Intell Med* 2019; 101: 101723.

[6] Fang G, Xu P, Liu W. Automated Ischemic Stroke Subtyping Based on Machine Learning Approach. *IEEE Access* 2020; 8: 118426–118432.

[7] Govindarajan P, Soundarapandian RK, Gandomi AH, et al. Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications* 2020; 32: 817–828.

[8] Ahmed H, Younis EMG, Ali AA. Predicting Diabetes using Distributed Machine Learning based on Apache Spark*. *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*. Epub ahead of print 2020. DOI: 10.1109/itce48509.2020.9047795.