Lecture Notes in Networks and Systems 462

Samiksha Shukla
Xiao-Zhi Gao
Joseph Varghese Kureethara
Durgesh Mishra *Editors*

Data Science and Security

Proceedings of IDSCS 2022

Springer

A Hybrid Feature Selection for Improving Prediction Performance with a Brain Stroke Case Study

D. Ushasree, A. V. Praveen Krishna, Ch. Mallikarjuna Rao, and D. V. Lalita Parameswari

Abstract In the contemporary era, artificial intelligence (AI) is making strides into every conceivable field. With advancements in place, there have been applications of machine learning (ML) in healthcare domain. Particularly for diagnosis of diseases with data-driven approach, ML algorithms are capable of learning from training data and make predictions. Many supervised ML algorithms came into existence with varied capabilities. However, they do rely on quality of training data. Unless quality of training data is ensured, they tend to result in mediocre performance. To overcome this problem, feature engineering or feature selection methods came into existence. From the literature, it is understood that feature selection plays crucial role in improving performance of prediction models. In this paper, a hybrid feature selection algorithm is proposed to leverage performance of machine learning models in brain stroke detection. The algorithm is named as Hybrid Measures Approach for Feature Engineering (HMA-FE). It returns best features that could contribute toward prediction of class labels. A prototype application is built to demonstrate the utility of the proposed framework and the underlying algorithms. The performance of prediction models are evaluated without and with feature engineering. Its empirical results showed the significant impact of proposed feature engineering on various brain stroke prediction models. The proposed framework adds value to Clinical Decision Support System (CDSS) used in healthcare units by supporting brain stroke diagnosis.

Keywords Feature selection · Brain stroke detection · Machine learning models · Classification

D. Ushasree (✉) · A. V. P. Krishna
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
e-mail: dupakuntlausha@gmail.com

A. V. P. Krishna
e-mail: praveenkrishna@kluniversity.in

Ch. M. Rao
Department of CSE, GRIET, Hyderabad, India

D. V. L. Parameswari
Department of CSE, GNITS, Hyderabad, India

S. Shukla et al. (eds.), *Data Science and Security*, Lecture Notes in Networks and Systems 462, https://doi.org/10.1007/978-981-19-2211-4_33

1 Introduction

Advancements in artificial intelligence (AI)-based approaches have paved way for improving quality of diagnosis in healthcare domain. Data-driven approaches that consider the patients' vitals could be used in clinical decision support systems (CDSS). Especially supervised machine learning models are widely used to detect various diseases in the healthcare industry. Brain stroke detection is one such application of machine learning (ML) algorithms. An important advantage of ML algorithms is that they can exploit historical data and the ground truth underlying in the data. They also suffer from mediocre performance when the training quality of data is inadequate. Due to the redundant and irrelevant features, the algorithms take more time and suffer from performance deterioration. In order to overcome these issues, feature engineering or feature selection approaches came into existence. They are broadly classified into three categories known as filter approaches, wrapper approaches and hybrid approaches. Filter approaches are based on relevance of features by correlating them with a dependent variable. On the other hand, wrapper approaches are based on finding usefulness of features by applying a training model. The former is much faster than the latter. In this paper a hybrid measures based method is followed that comes under filter approaches.

Many researchers contributed toward feature engineering. They defined different approaches or measures to determine best features. Liu et al. proposed a hybrid feature selection that combines phenotypic features and image features. The feature selection method is meant for improving prediction model for neuropsychiatric disorders. Leamy et al. focused on stroke detection and also studied EEG features and recovery of patients when neurorehabilitation therapy which is BCI-mediated. Kuhn et al. explored predictive models and the importance of feature engineering. Tasmin et al. explored different feature engineering methods such as tree-based feature selection, Random Forest, extra tree classifier, feature set generation and classifier-based models.

1.1 Problem Definition

Many supervised ML algorithms came into existence with varied capabilities. However, they do rely on quality of training data. Unless quality of training data is ensured, they tend to result in mediocre performance. To overcome this problem, feature engineering or feature selection methods came into existence. From the literature, it is understood that feature selection plays crucial role in improving performance of prediction models.

1.2 Motivation

Feature engineering with a hybrid approach could leverage brain stroke prediction performance. This will have impact on AI-based CDSSs in the real world applications used in healthcare industry. When detection accuracy is improved it will add to Quality of Service (QoS) in healthcare units. It is the motivation behind this research.

1.3 Contribution

In this paper, a hybrid feature selection algorithm is proposed to leverage performance of machine learning models in brain stroke detection. In this paper our contributions are as follows.

1. An algorithm known as Hybrid Measures Approach for Feature Engineering (HMA-FE) is defined based on a hybrid measure to identify importance of features. It gives the features that efficiently contribute the prediction of class labels toward brain stroke detection.
2. To demonstrate a prototype application is built for utility of the proposed framework and the underlying algorithms. The performance of prediction models is evaluated with and without feature engineering.

1.4 Organization of the Paper

The remainder of the paper is structured as follows. Section 2 reviews literature on different aspects of machine learning for brain stroke detection. Section 3 presents the proposed framework and underlying algorithm for efficient brain stroke detection. Section 4 presents performance evaluation. Section 5 gives paper conclusion and its directions for the future work.

2 Related Work

Literature reviews were given in this section on different brain stroke methods and feature selection approaches. Liu et al. [1] proposed a hybrid feature selection that combines phenotypic features and image features. The feature selection method is meant for improving prediction model for neuropsychiatric disorders. Katz et al. [2] proposed a methodology for comprehending the scale of prehospital stroke severity. Leamy et al. [3] focused on stroke detection and also studied EEG features and recovery of patients when neurorehabilitation therapy which is BCI-mediated. In terms of recovery of lost motor control, their research could help in improving patient

recovery. Vetten et al. [4] investigated on side effects associated with stroke. Particularly they studied on "acute corticospinal tract wallerian degeneration" results in poor motor outcome. Kuhn et al. [5] explored predictive models and the importance of feature engineering. Pathanjali et al. [6] studied different machine learning methods for ischemic stroke detection. Buck et al. [7] explored on ischemic stroke detection methods and further investigated on the relation between brain stroke and Neutrophilia development in patients. Kamel et al. [8] studied the after brain stroke effects of patients particularly on cardiac monitoring. They did it in order to identify atrial fibrillation in such patients and analyzed the cost effectiveness of their method. West et al. [9] focused on the Cryptogenic Stroke and the frequency of migraine and patent foramen ovale in patients. Soltanpour et al. [10] proposed a methodology to have automatic segmentation of ischemic stroke lesion with the help of CT perfusion maps. They used a deep learning model named MultiRes U-Net for this purpose.

Tasmin et al. [11] explored different feature engineering methods such as tree-based feature selection, Random Forest, extra tree classifier, feature set generation and classifier-based models. Lazer et al. [12] investigated on the re-emergence of stroke deficits with respect to Midazolam challenge. Tsivgoulis et al. [13] proposed a method to understand the mechanisms to prevent stroke second time by using cardiac rhythm monitoring. Parsons et al. [14] opined that Thrombolysis is one of the approaches to mitigate effects of stroke. In order to validate their study, they used "diffusion and perfusion weighted MRI". From the literature, it is understood that feature selection plays crucial role in improving performance of prediction models. In this paper, a hybrid feature selection algorithm is proposed to leverage performance of machine learning models in brain stroke detection.

3 Proposed Framework

The proposed methodology for brain stroke prediction is shown in Fig. 1. It has different mechanisms and underlying algorithms for brain stroke detection. The framework is aimed at having functional flow that takes brain stroke dataset as input and detects stroke probability of patients. The algorithms are used in order to have efficient detection of stroke with data-driven approach using supervised machine learning techniques.

The brain stroke detection dataset is subjected to preprocessing where the data is split into training set (80%) and testing set (20%). In the testing set, the class label is removed and used as ground truth. It is done so as the prediction models need to predict the class label. From the training data, if all the features are used for learning, it may lead to deteriorated performance due to redundant and irrelevant features. In order to overcome this and improve the efficiency of feature selection or feature engineering, an algorithm named Hybrid Measures Approach for Feature Engineering (HMA-FE) which makes use of two measures in combination. They are known as entropy and information gain. Entropy is a measure which finds uncertainty which is related to a given random variable while information gain computes the

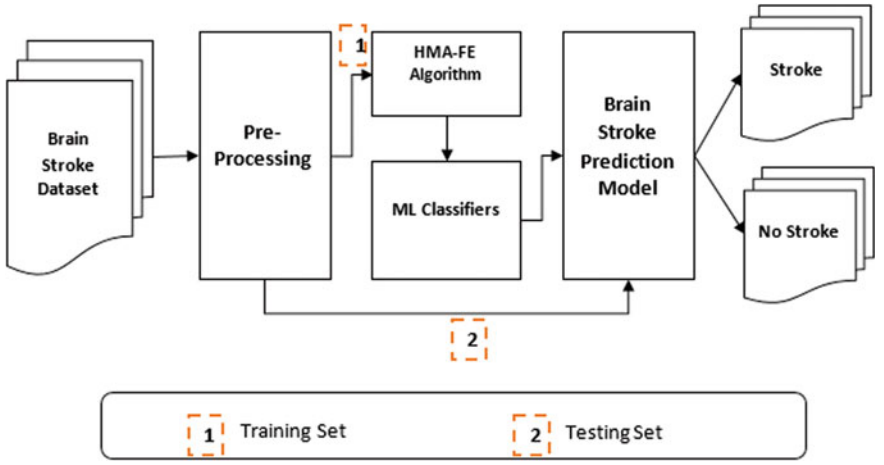

Fig. 1 Methodology for stroke prediction

amount of change in entropy.

$$H(X) = -\sum_{x \in X} p(x) \log \log p(x) \tag{1}$$

$$H(y) = -\sum_{y \in Y} p(y) \log \log p(y) \tag{2}$$

As shown in Eqs. 1 and 2 both $H(X)$ and $H(Y)$ are associated with the entropy measure. They are used to find information gain as in Eq. 3.

$$\text{Information gain} = H(y) - -H(y/x) \tag{3}$$

There is a hybrid measure known as symmetric uncertainty that combines both entropy and gain as in Eq. 4.

$$\text{SU} = \frac{2 * \text{Gain}}{H(x) + H(y)} \tag{4}$$

This measure is finally used to know the importance of a feature in the given brain stroke dataset. Since the feature engineering involves in finding importance of different features and choosing best contributing features, this measure assumes significance. It is used in the proposed algorithm named Hybrid Measures Approach for Feature Engineering (HMA-FE). An algorithm named Hybrid

Measures Approach for Feature Engineering (HMA-FE) is proposed and implemented for identification of best features that contribute to the prediction of brain stroke.

Algorithm 1: Hybrid Measures Approach for Feature Engineering

Algorithm: Hybrid Measures Approach for Feature Engineering (HMA-FE)

Input: Brain stroke dataset ***D***, importance threshold ***th***

Output: Selected features ***F***
1. Start
2. Initialize a map for holding su ***M***
Extract All Features
3. ***F***<GetAllFeatures(***D***)
Compute Importance Measure
4. For each ***f*** in ***F***
5. ***e***<FindEntropy(***f***,***F***) //using Eqs. 1 and 2
6. ***ig***<FindInforGain(***f***,***F***) // using Eq. 3
7. ***su***<FindSU(***e***, ***ig***) // using Eq. 4
8. ***M***<Add(***f***, ***su***)
9. End For
Find Contributing Features
10. For each ***f*** in ***F***
11. ***su***<GetFromMap(***f***,***M***)
12. IF ***su<th*** THEN
13. Remove ***f*** from ***F***
14. End If
15. End For
Return Contributing Features
16. Return ***F***
17. End

As presented in Algorithm 1, a hybrid measures approach is defined to have feature engineering. The algorithm takes brain stroke dataset denoted as *D* and an importance threshold th as input and finds contributing features that are useful for efficient prediction of brain stroke. Step 2 initializes a map denoted as *M* for holding intermediary outcomes in the form of features and symmetric uncertainty values as key/value pairs. Step 3 g*et all* features from *D*. Step 4 through Step 9 an iterative process computes symmetric uncertainty from measures such as entropy and information gain and *M* gets updated with corresponding feature and its symmetric uncertain value. After completion of this iterative process, there is need for finding contributing features or the features that are very useful in class label prediction. Step 10 through Step 15 encapsulate an iterative process that identifies contributing features based on a threshold value. The features that exhibit symmetric uncertainty less than the threshold value are discarded. After completion of this process, only the contributing features remain in the vector *F*. Such features are finally returned by the algorithm.

4 Dataset and Experimental Setup

Dataset is collected from [15] where it has 11 features of clinical importance for brain stroke detection. The attributes include patient id, gender, age, hypertension, heart disease, ever married, residence type, average glucose type, BMI and smoking status. Trials were made with a prototype built using Data Science (Python) platform. Performance evaluation is made using standard metrics such as precision, recall, F1-score and accuracy.

5 Results and Discussion

This section provides performance of different prediction models in terms of precision, recall, F1-score and accuracy.

As presented in Fig. 2, the precision and recall performance of different brain stroke prediction models is provided. The performance metrics such as precision and recall are presented in horizontal axis and the values for precision and recall are shown in vertical axis. The value for precision and recall can be in the range between 0.0 and 1.0. More value indicates better performance. Many prediction models exhibited precision as 1.0 while Decision Tree classifier and KNeighbors classifier showed 0.55882 and 0.02941, respectively. Neural Nets classifier showed 0.05882 as precision. With respect to recall, highest recall, 1.0, is exhibited by KNeighbors classifier. Gaussian NB showed 0.04985, Decision Tree classifier 0.05 and Neural Nets 0.05882. Interestingly many prediction models showed same recall 0.96. They include Bernoulli NB, Random Forest, Logistic Regression, Gradient Boosting classifier, Support Vector Machine and Stochastic Gradient Descent.

As presented in Fig. 3, the F-measure and accuracy performance of different brain stroke prediction models is provided. The performance metrics are presented in horizontal axis and the vertical axis shows the values for F-measure and accuracy. The value for F-measure and accuracy can be in the range between 0.0 and

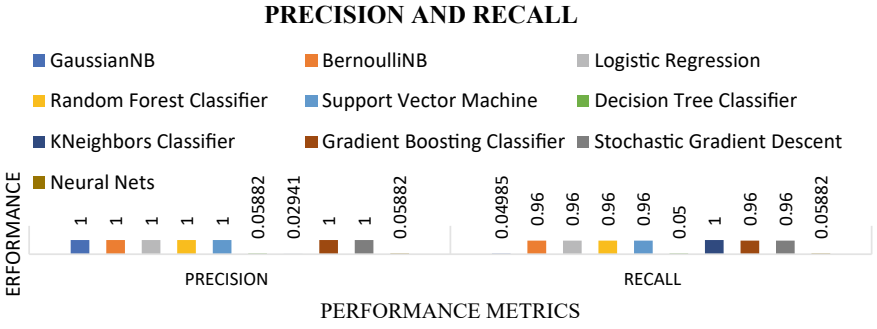

Fig. 2 Performance assessment in terms of precision and recall

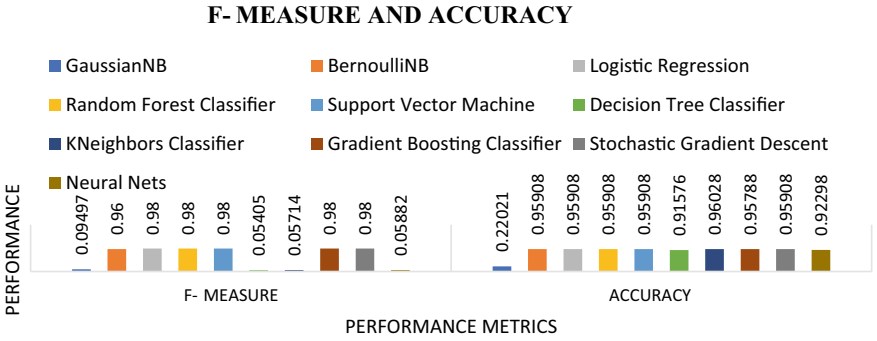

Fig. 3 Performance comparison in terms of F-measure and accuracy

1.0. More value indicates better performance. Different prediction models showed varied performance. The highest F-measure is exhibited by several prediction models like Logistic Regression, Random Forest, SVM, Gradient Boosting classifier and Stochastic Gradient Boosting. Highest accuracy is exhibited by KNeighbors Classifier with 0.96028 and the least performance is shown by GaussianNB with 0.22021. The second highest performance in terms of accuracy with 0.95908 is shown by various prediction models like BernoulliNB, Logistic Regression, Random Forest Classifier, Support Vector Machine and Stochastic Gradient Descent. Other prediction models such as Decision Tree Classifier showed 0.91576, KNeighbors Classifier 0.96028, Gradient Boosting Classifier 0.95788 and Neural Nets 0.92298.

As presented in Fig. 4, the performance of the prediction models is higher when feature selection algorithm is used. Brain stroke prediction models are provided in horizontal axis while accuracy of the models is shown in vertical axis. Highest accuracy is exhibited by KNeighbors Classifier with 0.96028 and the least performance is shown by GaussianNB with 0.22021. The second highest performance in terms of accuracy with 0.95908 is shown by many prediction models such as Logistic

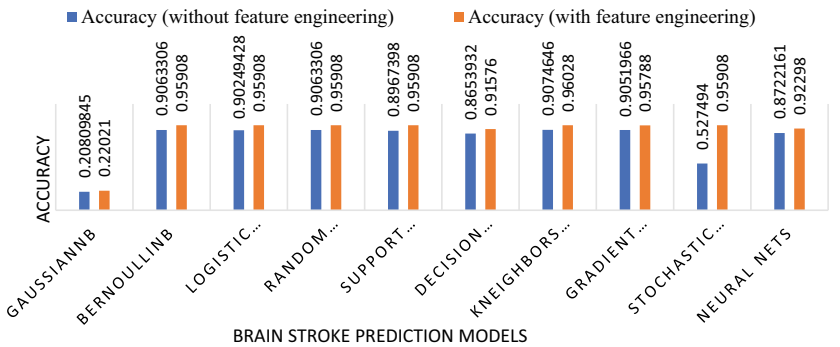

Fig. 4 Performance comparison of prediction models with and without feature selection

Regression, BernoulliNB, Random Forest Classifier, Support Vector Machine and Stochastic Gradient Descent. Other prediction models like Decision Tree Classifier showed 0.91576, KNeighbors Classifier 0.96028, Gradient Boosting Classifier 0.95788 and Neural Nets 0.92298. All the models showed less performance consistent when the proposed feature selection algorithm is not used.

6 Conclusion and Future Work

In this paper, a feature selection algorithm known as Hybrid Measures Approach for Feature Engineering (HMA-FE) is proposed and implemented. The algorithm is based on a combined metric that exploits entropy and information gain measures to have better probability of identifying features that contribute to the brain stroke prediction. Different machine learning models are used for brain stroke prediction using supervised learning approach. The models include GaussianNB, BernoulliNB, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Tree Classifier, KNeighbors Classifier, Gradient Boosting Classifier, Stochastic Gradient Descent and Neural Nets. These prediction models showed improved performance in terms of accuracy when the proposed feature engineering method is used. Data science platform using Python is used to implement the proposed algorithm and to evaluate the prediction models. The empirical results reveal the performance improvement when quality of training is improved using the proposed feature selection method. In future, we intend to improve brain stroke prediction models using ensample approach.

References

1. Liu L, Tang S, Wu F, Wang Y-P, Wang J (2021) An ensemble hybrid feature selection method for neuropsychiatric disorder classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1
2. Katz BS, McMullan JT, Sucharew H, Adeoye O, Broderick JP (2015) Design and validation of a prehospital scale to predict stroke severity. Stroke 46(6):1508–1512
3. Leamy DJ, Kocijan J, Domijan K, Duffin J, Roche RAP, Commins S, Collins R, Ward TE (2014) An exploration of EEG features during recovery following stroke – implications for BCI-mediated neurorehabilitation therapy. J Neuroengineering Rehabilitation, 1–16
4. DeVetten G, Coutts SB, Hill MD, Goyal M, Eesa M, O'Brien B, Demchuk AM, Kirton A (2010) Acute corticospinal tract wallerian degeneration is associated with stroke outcome. Stroke 41(4):751–756
5. Butcher B, Smith BJ (2020) Feature engineering and selection: A practical approach for predictive models. Am Stat 74(3):308–309
6. Pathanjali C, Priya T, Monisha G, Bhaskar S (2020) Machine learning for predicting ischemic stroke. IJERT 9(5):1–4
7. Buck BH, Liebeskind DS, Saver JL, Bang OY, Yun SW, Starkman S, Ali LK, Kim D, Villablanca JP, Salamon N, Razinia T, Ovbiagele B (2008) Early neutrophilia is associated with volume of ischemic tissue in acute stroke. Stroke 39(2):355–360

8. Kamel H, Hegde M, Johnson DR, Gage BF, Johnston SC (2010) Cost-effectiveness of outpatient cardiac monitoring to detect atrial fibrillation after ischemic stroke. Stroke 41(7):1514–1520
9. West BH, Noureddin N, Mamzhi Y Low CG, Coluzzi AC, Shih EJ, Gevorgyan Fleming R, Saver JL, Liebeskind DS, Charles A, Tobis JM (2018) Frequency of patent foramen ovale and migraine in patients with cryptogenic stroke. Stroke, 1–7
10. Soltanpour M, Greiner R, Boulanger P, Buck B (2021) Improvement of automatic ischemic stroke lesion segmentation in CT perfusion maps using a learned deep neural network. Comput Biol Med 137:104849
11. Tasmin M, Ishtiak T, Uddin Ruman S, Ur Rahaman Chowdhury Suhan A, Shihab Islam NM, Jahan S, Ahmed S, Shahnawaz Zulminan Md, Raufus Saleheen A, Rahman RM (2020). [IEEE 2020 IEEE 10th International Conference on Intelligent Systems (IS) - Varna, Bulgaria (2020.8.28–2020.8.30)] 2020 IEEE 10th International Conference on Intelligent Systems (IS)—Comparative Study of Classifiers on Human Activity Recognition by Different Feature Engineering Techniques, 93–101
12. Lazar RM, Fitzsimmons B-F, Marshall RS, Berman MF, Bustillo MA, Young WL, Mohr JP, Shah J, Robinson JV (2002) Reemergence of stroke deficits with midazolam challenge. Stroke 33(1):283–285
13. Tsivgoulis G, Katsanos AH Grory BM, Köhrmann M, Ricci BA, Tsioufis K, Cutting S, Krogias C, Schellinger PD, Campello AR, Cuadrado-Godia E, Gladstone DJ, Sanna T, Wachter R, Furie K, Alexandrov AV, Yaghi S (2019) Prolonged cardiac rhythm monitoring and secondary stroke prevention in patients with cryptogenic cerebral ischemia. Stroke, 1–6
14. Parsons MW, Alan Barber P, Chalk J, Darby DG, Rose S, Desmond PM, Gerraty RP, Tress BM, Wright PM, Donnan GA, Davis SM (2002). Diffusion- and perfusion-weighted MRI response to thrombolysis in stroke. 51(1):28–37
15. Comparison of 12 risk stratification schemes to predict stroke in patients with nonvalvular atrial fibrillation. Stroke 39(6):1901–1910