Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

A Machine Learning Framework with Hybrid Feature Engineering for Leveraging Brain Stroke Detection Performance

Kashi Sai Prasad [1], Dr. S. Pasupathy [2]

[1] Research Scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamilnadu.

[2] Associate Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamilnadu.

Article Info
Page Number: 1805 – 1824
Publication Issue:
Vol. 71 No. 3s2 (2022)

Article History
Article Received: 22 April 2022
Revised: 10 May 2022
Accepted: 15 June 2022
Publication: 19 July 2022

Abstract

Machine Learning (ML) models are widely used in solving real world problems in all domains and healthcare is no exception. Since supervised learning methods in machine learning exploit learning from data to gain Artificial Intelligence (AI), it is indispensable to have adequate training sample that are of high quality. If not the performance of ML models will be deteriorated. In order to overcome this problem and enhance the state of the art in supervised learning, in this paper, we proposed a ML framework known as Brain Stroke Detection Framework (BSDF). We also proposed a hybrid feature engineering method that will be used in ML pipeline of the framework for leveraging prediction performance. The algorithm is known as Hybrid Feature Engineering (HFE) which is the combination of three filter based approaches. The framework is realized with another algorithm known as Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD) which exploits HFE for improving prediction performance. It is a data driven approach to have cheaper alternative to complement Clinical Decision Support System (CDSS) in healthcare units. Many brain stroke prediction models could achieve 97% accuracy when HFE is used as underlying feature selection method. There is significant improvement in performance of different brain stroke prediction models with the hybrid feature engineering algorithm.

Keywords: Machine Learning, Feature Engineering, Brain Stroke Detection, Supervised Learning, Hybrid Feature Engineering.

1. Introduction

Data in every domain is given high importance due to the availability of computational intelligence methods. Artificial Intelligence (AI) based approaches with machine learning paved way for exploiting data to the fuller extent. With the emergence of cloud computing and big data ecosystem and distributed computing platforms, data is given unprecedented importance by enterprises in the real world. Data is any domain helps in acquiring business intelligence (BI). Similarly, data is healthcare domain is given more importance due to the possibilities to improve Quality of Service (QoS) in healthcare units. In this paper, brain stroke detection research is carried out due to the increasing mortality and disability rate across the globe due to brain stroke as per WHO statistics. Many ML approaches are used by the researchers and there is vast literature found on healthcare research exploiting ML models.

Badriyah *et al*. [10] focused on stroke classification with many ML algorithms. It includes image processing and feature extraction procedure. Sugnaya and Murugavalli [13] proposed a

feature selection method that is used for ML based detection of a lingual script. Açikoğlu and Tuncer [14] used feature selection based approach along with ML techniques for neonatal seizure diagnosis. Hung *et al*. [15] used ML and deep learning methods for stroke prediction. Ray *et al.* [16] proposed a feature selection method for brain stroke detection using Chi-Square method. Suresh and Duerstock [18] proposed an optimal feature selection method for detecting diseases from given sample data. There are many contributions in the literature on usage of ML for brain stroke detection as studied in [2], [3], [6], [9] and [11]. However, feature selection is found essential to leverage prediction performance.

From the literature it is understood that ML models are widely used for brain stroke detection. However, there is problem with low quality feature space and therefore, it is still an active research area to reduce dimensionality in the given dataset with more efficient feature selection approaches. To overcome this problem and enhance the state of the art in supervised learning, in this paper, we proposed a ML framework known as Brain Stroke Detection Framework (BSDF). We also proposed a hybrid feature engineering method that will be used in ML pipeline of the framework for leveraging prediction performance. Our contributions in this paper are as follows.

1. We proposed a ML based framework known as Brain Stroke Detection Framework (BSDF) with multi-model pipeline for brain stroke detection.
2. We proposed a hybrid feature engineering method that is used in ML pipeline of the framework for leveraging prediction performance. The algorithm is known as Hybrid Feature Engineering (HFE) which is the combination of three filter based approaches.
3. The framework is realized with another algorithm known as Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD) which exploits HFE for improving prediction performance.
4. A prototype application is built using Python data science platform to evaluate the proposed framework and underlying algorithms with benchmark dataset and the dataset collected from a corporate healthcare unit.

The remainder of the paper is structured as follows. Section 2 reviews literature on brain stroke detection and feature selection methods. Section 3 presents the proposed methodology, framework, algorithms and evaluation methodology. Section 4 presents experimental results and evaluation of the same. Section 5 concludes the paper and gives possible directions for future work.

2. Related Work

This section reviews literature on various models used for brain stroke detection and it also throws light on feature selection methods. Sirsat *et al*. [1] made a systematic review of literature on brain stroke detection using ML approaches. They found that brain stroke has second highest mortality rate. They studied ML methods used for stroke detection, stroke diagnosis and stroke prognosis besides treatment. Emon *et al*. [2] explored many ML methods such as "Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier,

Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer Perceptron Classifier, KNeighbors Classifier, Gradient Boosting Classifier, and XGBoost Classifier" for brain stroke detection. Pradeepa *et al*. [3] used social media data to investigate with ML techniques on the discovery of risk factors of stroke. They used Twitter API to collect corpus and used Natural Language Processing (NLP) and ML for predicting risk factors. Kamal *et al*. [4] explored machine learning techniques on acute ischemic stroke. Sample size and labelling of regions in brain neuroimaging are the problems identified by them.

Mohanty *et al*. [5] focused on rehabilitation research on brain stroke patients. They discussed about Brain Computer Interface (BCI) therapy and functional connectivity to do research on stroke rehabilitation. Govindarajan *et al*. [6] proposed and implemented a prototype for brain stroke detection using ML. They worked on the raw data collected from the healthcare units. SVM, decision tree and logistic regression are used for classification. Salucci *et al*. [7] investigated on learning by examples technique with machine learning in order to find stroke in real time. In the process, they used SVM technique in order to predict brain stroke probability. Lee *et al*. [8] proposed a methodology to detect brain stroke probability within 4.5 hours using machine learning and image processing. Choi *et al*. [9] on the other hand proposed an elderly stroke monitoring system using ML models and vital signals of Electroencephalography. Badriyah *et al*. [10] focused on stroke classification with many ML algorithms. It includes image processing and feature extraction procedure.

Islam *et al*. [11] focused on stroke prediction research using ML algorithms like Random Forest. They found that RF is better than other techniques like K-NN, decision tree and logistic regression. Fang *et al*. [12] used many ML techniques such as "linear SVC, Random-Forest-Classifier, Extra-Trees-Classifier, AdaBoost-Classifier, and Multinomial-Naïve-Bayes-Classifier" along with cross validation to have better performance in brain stroke detection. Sugnaya and Murugavalli [13] proposed a feature selection method that is used for ML based detection of a lingual script. Açikoğlu and Tuncer [14] used feature selection based approach along with ML techniques for neonatal seizure diagnosis. Hung *et al*. [15] used ML and deep learning methods for stroke prediction. Ray *et al*. [16] proposed a feature selection method for brain stroke detection using Chi-Square method. Salucci *et al*. [17] investigated on biomedical imaging with learning by examples in order to detect diseases. Suresh and Duerstock [18] proposed an optimal feature selection method for detecting diseases from given sample data. Ang *et al*. [19] also explored BCI research in order to ascertain discriminatory features leading to disease diagnosis. Yeh [20] proposed a healthcare system that involves body sensor networks using ML approaches. From the literature it is understood that ML models are widely used for brain stroke detection. However, there is problem with low quality feature space and therefore, it is still an active research area to reduce dimensionality in the given dataset with more efficient feature selection approaches.

3. Proposed Methodology

The proposed methodology for brain stroke detection is based on supervised learning approach coupled with a hybrid feature selection method. It is a data driven approach to have cheaper alternative to complement Clinical Decision Support System (CDSS) in healthcare units. Since

supervised learning methods in machine learning exploit learning from data to gain Artificial Intelligence (AI), it is indispensable to have adequate training sample that are of high quality. If not the performance of ML models will be deteriorated. In order to overcome this problem and enhance the state of the art in supervised learning, in this paper, we proposed a hybrid feature engineering method that will be used in ML pipeline for leveraging prediction performance. More details are provided in the ensuing sub sections.

3.1. The Framework

We proposed a framework named as Brain Stroke Detection Framework (BSDF) which is used to have complete mechanisms required to automate the brain stroke detection process. It has components that are reusable make workflow towards brain stroke detection with improved efficiency. As explored in [1], [3], [5], [6], ML models are widely used for solving real world problems. With the availability of training data, ML models are growing in popularity in every conceivable domain including healthcare industry. At the same time, the performance of the ML models is severely affected when training data is of less quality. It happens that the data used for training might have inconsistencies and inherent quality issues like redundancy, curse of dimensionality and irrelevant features. Unless the feature space is reduced by eliminating irrelevant and redundant features, the ML models exhibit mediocre performance. This fact is evident in the research found in [11], [12], [14] and [16]. In order to overcome this problem and leverage state of the art in feature engineering, we proposed an algorithm known as Hybrid Feature Engineering (HFE).

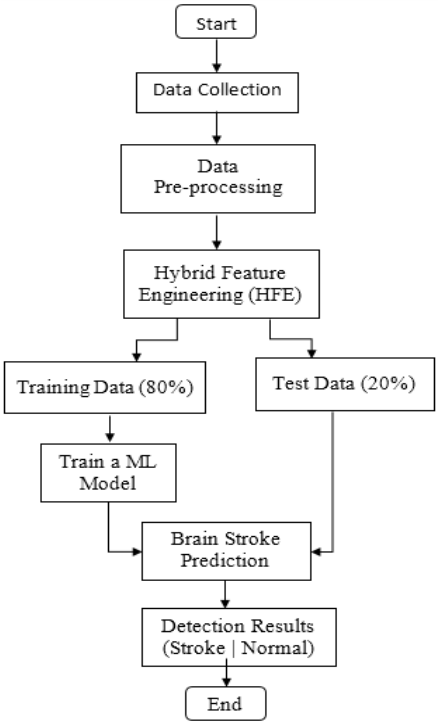

Figure 1; Proposed Brain Stroke Detection Framework (BSDF)

As presented in Figure 1, the proposed framework BSDF has provision to exploit feature selection prior to learning from the training dataset. After collection of data based on the

procedure discussed in Section 3.2, the data is subjected to pre-processing where it is divided into 80% training data and 20% test data. It is done as there needs to be sufficient training data. The proposed feature engineering algorithm named HFE, discussed in Section 3.3, is employed on both training and testing data in order to reduce feature space. Several ML models are used in pipeline to learn from the training data and make predictions on the test data. The performance of the prediction models is evaluated using the procedure described in Section 3.5.

3.2. Data Collection Procedure

Data for brain stroke detection research is collected from [21]. It has 5110 instance covering both normal and brain stroke probability patients. Each instance in the dataset is pertaining to a patient. It has 12 attributes and the last one is the ground truth containing diagnosis value or class label with 0 and 1 indicating NORML and STROKE respectively. The data includes many health parameters of patients that are crucial for stroke probability detection.

Table 1. Shows all the attributes and their values for clear understanding.

Table 1. Patient Attributes in the Dataset

Attribute Name	What it does?	Value Range
ID	Every patient is uniquely identified with the ID.	Any unique value.
GENDER	Holds the value of gender.	1. Male 2. Female 3. Other
AGE	Holds the age of a patient.	A range of value from 1 (or less than 1) to up to 100.
HYPERTENSION	Holds the hypertension status of the patient.	1. Value 0 indicates no hypertension 2. Value 1 indicates hypertension
HEART DISEASE	Holds the heart disease status of the patient.	1. Value 0 indicates no heart disease 2. Value 1 indicates heart disease
EVER MARRIED	Holds marital status of patient.	1. Yes 2. No
WORK TYPE	Holds the kind of work doing by patient.	1. Children 2. Govt Job 3. Never Worked 4. Private 5. Self Employed
RESIDENCE TYPE	Holds the kind of residence of patient.	1. Rural 2. Urban

AVG_GLUCOSE_LEVEL	Holds the value reflecting average glucose level of patient.	A numeric value reflecting average glucose value.
BMI	Holds BMI value of patient	A numeric value reflecting BMI value.
SMOKING	Holds smoking status of patient	1. Formerly smoked 2. Never smoked 3. Smokes 4. Unknown
STROKE	Class label attribute	1. Value 1 indicates stroke 2. Value 0 indicates normal

The data collected from [21] contains attributes as provided in the table. Keeping the attributes remain same, the researcher has added more authentic instances to the dataset. Additional data is collected from Malla Reddy Narayana Multispecialty Hospital [22] located in Hyderabad. Totally 2112 new patients' data is added to the dataset. The total number of records thus reached to 7222.

3.2. Data Splitting

The dataset acquired is divided into 80% training data and 20% testing data. An excerpt of 10 instances collected from training data is provided in Table 2. Observe the last attribute is *stroke* which indicates class label. This attribute does not exist in the test data as it has to be predicted by the ML models.

Table 2. Shows an Excerpt from the Training Data

Id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1

37089	Female	37	1	0	Yes	Self-employed	Rural	127.71	36	never smoked	0
68614	Female	48	0	0	Yes	Private	Rural	216.7	38.7	formerly smoked	0
1686	Female	29	0	0	No	Private	Urban	71.89	27.6	never smoked	0
22284	Male	22	0	0	No	Private	Rural	103.56	25.1	Unknown	0
39038	Male	11	0	0	No	Children	Rural	79.03	16.5	Unknown	0

As presented in Table 2, an excerpt of data collected from the training set is provided to ascertain the attributes and values.

3.3. Feature Engineering

Feature engineering or feature selection is used to reduce dimensions in the data so as to leverage prediction performance. Often it solves the problem known as curse of dimensionality. There are two broad approches tofeature selection. They are known as filter based approches and wrapper based approaches. We proposed an algorithm named Hybrid Feature Engineering (HFE) based on filter based approach which essentially looks like the process shown in Figure 2.

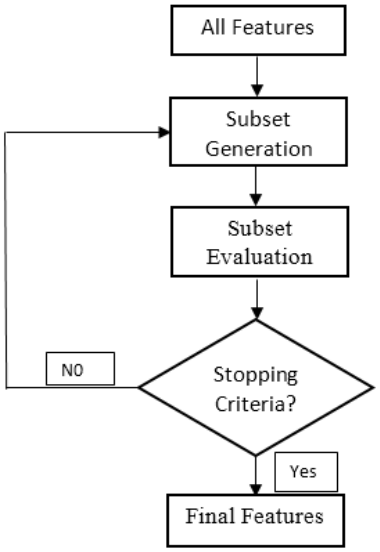

Figure 2; A Typical Filter based Approach to Feature Selection

The filter based method is based on certain metric to fine the feature importance. It takes all features and performs subset generation and evaluates the feature subset. This is an iterative process until stopping criteria is satisfied. Finally, it returns all selected features. Based on this filter based approach, we proposed a feature selection method known as Hybrid Feature Engineering (HFE). It is a hybrid of three filter methods known as Fisher criterion, entropy and t-test.

Table 3. Notations Used in the Proposed Hybrid Feature Engineering

Notation	Description
$\mu_1(i)$ and $\sigma_i(i)$	Mean value
n_1 and n_0	The number of patterns in the null and unitary class
KL-distance	Kullback Liebler distance
P	Probability distribution
Q	Target probability distribution

Table 3 shows the notations used in the proposed feature engineering algorithm. Fisher index computation as explored in [23] is widely used for feature selection. It is computed as in Eq. 1.

$$F(i) = |\frac{\mu_1(i)-\mu_0(i)}{\sigma_1^2(i)-\sigma_0^2(i)}| \quad (1)$$

Figure index provides importance of each variable or feature associated with the underlying dataset. T-test is another widely used filter approach as discussed in [24]. It is used to evaluate relative importance of each feature and it is computed as in Eq. 2.

$$t(i) = |\frac{\mu_1(i)-\mu_0(i)}{\sqrt{\frac{\sigma_1^2(i)}{n_1}+\frac{\sigma_0^2(i)}{n_0}}}| \quad (2)$$

Yet another widely used filter approach is relative entropy. It is also known as Kullback-Leibler divergence as discussed in [25]. It is a distance function between two probability distributions.

$$\text{KL(p, q)} = \sum_i p_i log_2(\frac{p_i}{q_i}) \quad (3)$$

These three measures are used in the proposed hybrid approach to reap benefits of them in choosing best features for brain stroke detection research.

Algorithm 1. Hybrid Feature Engineering Algorithm

Algorithm: Hybrid Feature Engineering (HFE)
Input: Patient dataset with electronic health records for training *D*, threshold *th*
Output: Selected features that contribute to brain stroke diagnosis *F*
1. Start
2. Initialise attributes vector *A*
3. Initialise features vector *F*
4. Initialize feature scores map *M*

Find All Attributes

5. *A*←GetAllAttributes(*D*)

Extract All Features

6. For each *a* in *A*
7. *features*←IdetifyFeatures(*a*)
8. *F*←*F*+*features*
9. End For

Hybrid Filter based Feature Selection

10. For each *f* in *F*
11. *fisher_score*←FindFisherScore(*f*, *F*); //use Eq. 1
12. *t_test_score*←FindTTestScore(*f*, *F*); //use Eq. 2
13. *rentropy_score*←FindREntropyScore(*f*, *F*); //use Eq. 3
14. *mean_score*←FindMeanScore(*fisher_score*, *t_test_score*, *erentropy_score*)
15. Add *f* and *mean_score* to M
16. End For

Final Selection of Features

17. *F*←Empty
18. For each entry *m*in *M*
19. IF *m.mean_score* satisfies*th* THEN
20. Add *m.f* to *F*
21. End If
22. End For
23. Return *F*
24. End

As presented in Algorithm 1, it takes Patient dataset with electronic health records for training *D*, threshold *th* as inputs and returns the selected features. The algorithm has different stages of execution. They are known as finding all attributes, extracting all features from all attributes to create complete feature space, application of hybrid filter based approach for feature selection and finally arriving at final selection of features.

3.4. Brain Stroke Prediction Models

We proposed an algorithm named Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD) that has a pipeline of ML models for brain stroke prediction and evaluation.

Algorithm 2. Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD)

Algorithm: Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD)
Inputs: Patient dataset D, pipeline of ML models M
Output: Predictions P

1. Start
2. Initialize confusion matrix map C
3. Initialize prediction results map R
4. Initialize features vector F

```
5.    (T1, T2)←DataSplitting(D)
6.    F←Invoke HFE(T1)
7.    For each model m in M
8.      Train the m with F
9.      m←FitModel(T2)
10.     Update C with confusion matrix and m
11.   Add m predictions to R
12.   End For
13.   For each r in R
14.     Display prediction results
15.   End For
16.   For each c in C
17.   Compute precision using Eq. 4
18.   Compute recall using Eq. 5
19.     Compute F1-measure using Eq. 6
20.     Compute accuracy using Eq. 7
21.   Display evaluation results
22.   End For
23.   End
```

The proposed algorithm SML-BSD takes inputs such as Patient Dataset D, pipeline of ML models M. In Step 5, data is subjected to splitting into 80% training T1 and 20% testing T2. Step 6 invokes Algorithm 1 that is HFE algorithm which does feature engineering and returns only selected features. This step is crucial to improve performance of the prediction models that are part of the pipeline used. Step 7 through Step 12, an iterative process involves in training of all the models in pipeline and provide prediction model for each technique. The prediction model is fit and T2 is used for predictions. The resultant confusion matrix of each model is saved to C. Step 13 through 15 have an iterative process to display prediction results for test data. Step 16 through 22, there is an iterative process used to compute different performance metrics from confusion matrices C. For each model it prints performance evaluation values.

3.5. Performance Evaluation Metrics

Performance evaluation metrics used in this paper are precision, recall, F1-measure and accuracy. These are widely used metrics by the researchers of machine learning. The basis for these metrics is the confusion matrix which helps in ascertaining prediction performance when compared with ground truth.

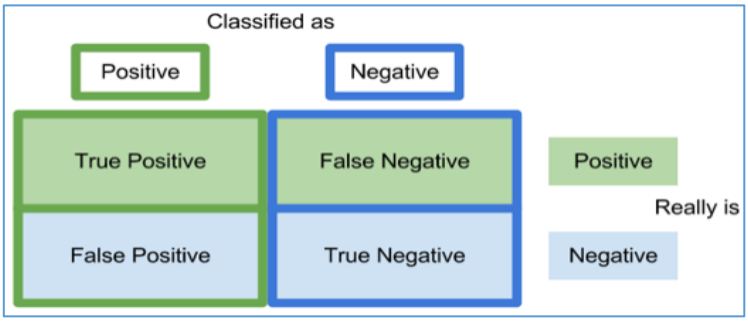

Figure 3; Confusion Matrix Model

From Figure 3, true positive (TP) refers to the fact that a patient has brain stroke and the ML model also predicted the same. False positive (FP) refers to the fact that a patient has no brain stroke and the ML model predicted stroke. Falsenegative (FN) refers to the fact that a patient has brain stroke and the ML model predicted no stroke. True negative (TN) refers to the fact that a patient has no brain stroke and the ML model also predicted no brain stroke. Based on these observations, precision, recall, F1-measure and accuracy are the metrics derived as expressed in Eq. 4, Eq. 5, Eq. 6 and Eq. 7 respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1-measure} = 2 * \frac{(precision * recall)}{(precision+recall)} \quad (6)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Each of these metrics can have a value from 0 to 1. The higher value indicates more performance in brain stroke prediction.

4. Results and Discussion

In this section, the experimental results are discussed and evaluated with different performance metrics. This section is further divided into three more sections covering exploratory data analysis that visualizes data distribution dynamics, feature importance which shows the feature ranking made by the proposed HFE algorithm and the evaluation of the results.

4.1. Exploratory Data Analysis

This section presents data dynamics associated with the dataset used for empirical study. The stroke probability relation of age and gender, age and marital status, age and smoking status and age and work type is analysed. Age distribution, average glucose level distribution and Body Mass Index (BMI) distribution are visualized with a histogram.

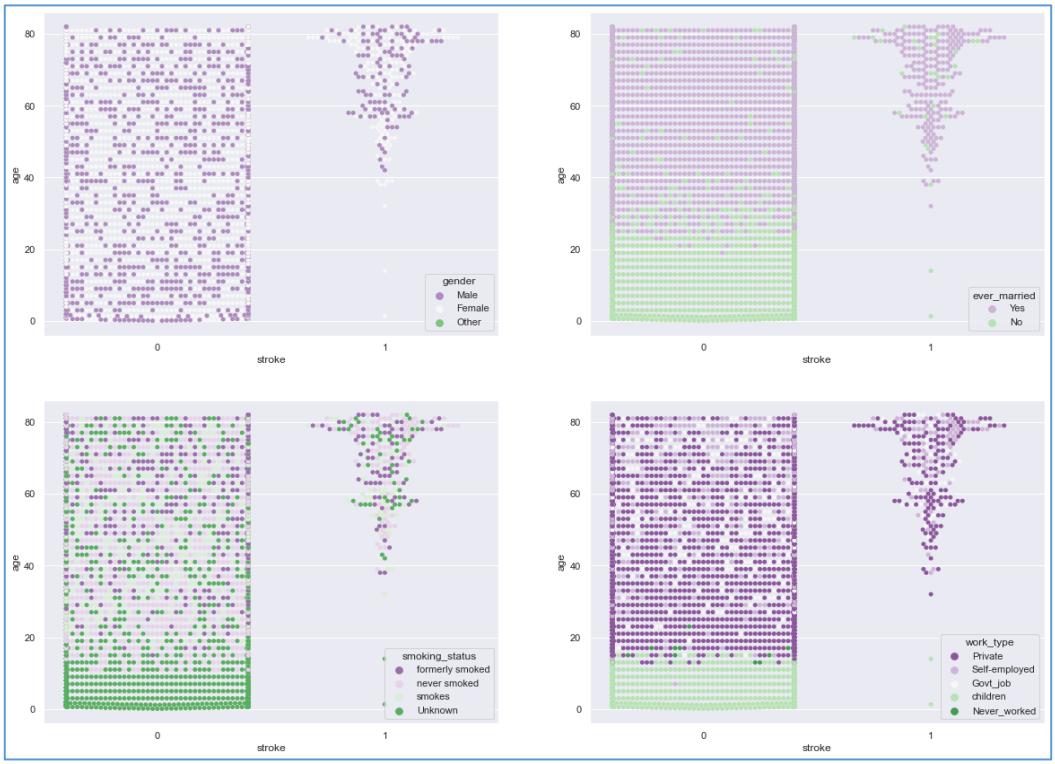

Figure 4; Stroke Probability Analysis with Gender, Marital Status, Smoking Status and BMI against Age

As presented in Figure 4, the brain stroke probability analysis with gender, marital status, smoking status and BMI against age is visualized. In the horizontal axis 0 and 1 are given indicating NORMAL and STROKE respectively. In the vertical axis age is provided. The relationship of attributes such as gender, marital status, smoking status and BMI against age is provided with respect to stroke probability.

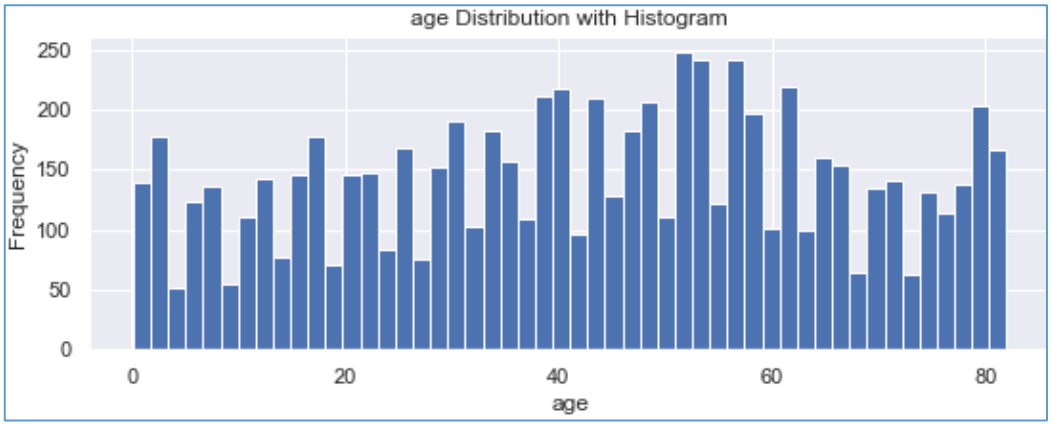

Figure 5; Age Distribution in the Stroke Dataset

As presented in Figure 5, the histogram visualization shows that right from below 1 year of age to 80 years age patients data is found in the dataset. The vertical axis shows the count of samples for each age in dataset.

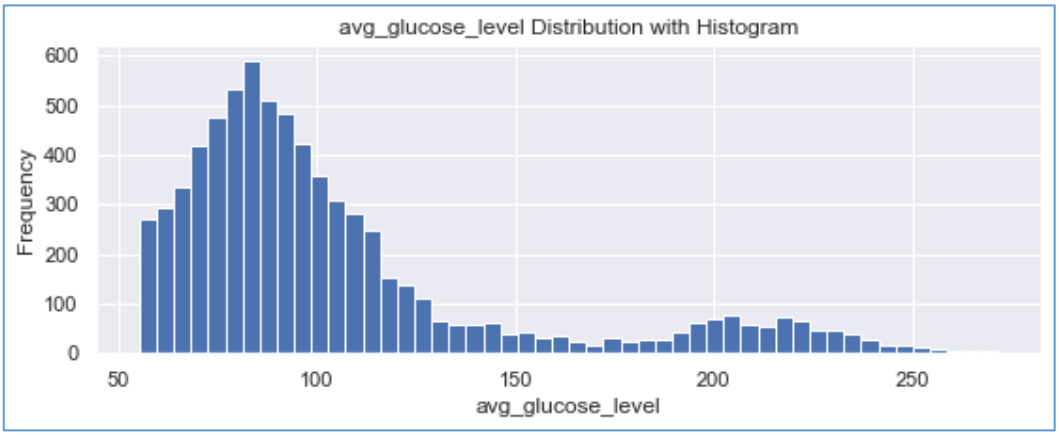

Figure 6; Average Glucose Distribution in the Stroke Dataset

As presented in Figure 6, the histogram visualization shows that right from above 50 to more than 150 average glucose level patients are found in the dataset. The vertical axis shows the count of samples against each average glucose level.

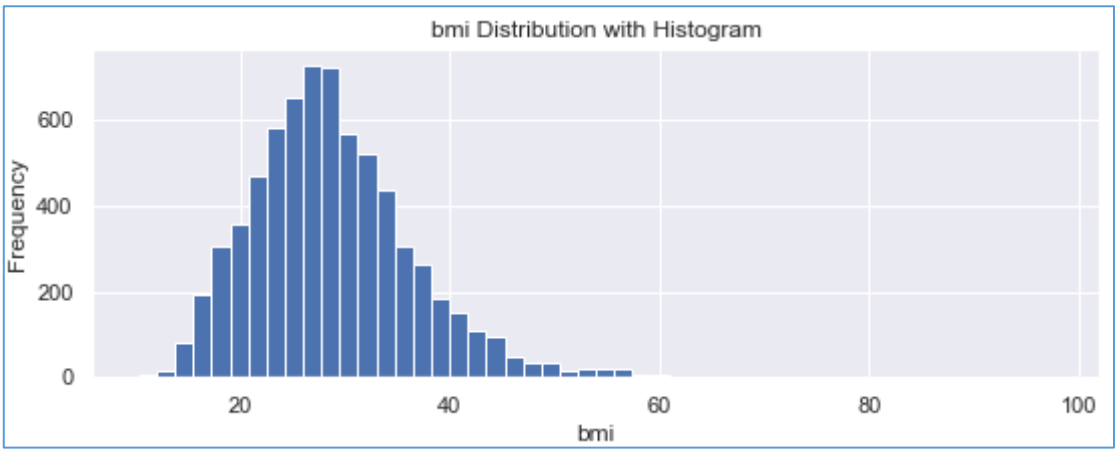

Figure 7; BMI Distribution in the Stroke Dataset

BMI has its related with stroke disease. It is a numerical value derived from body weight associated with height. In general, normal weight is indicated by BMI value between 18.5 and 25 kg/m². Underweight is denoted by BMI values less than 18.5 kg/m² and obese is reflected by BMI higher than 30 kg/m². As presented in Figure 7, the histogram visualization shows the BMI of patients in horizontal axis and corresponding frequency of samples in the vertical axis.

4.2. Feature Importance

The proposed HFE algorithm is used for finding importance of features. Identifying the features that contribute to the prediction of brain stroke significantly is the phenomenon known as feature engineering.

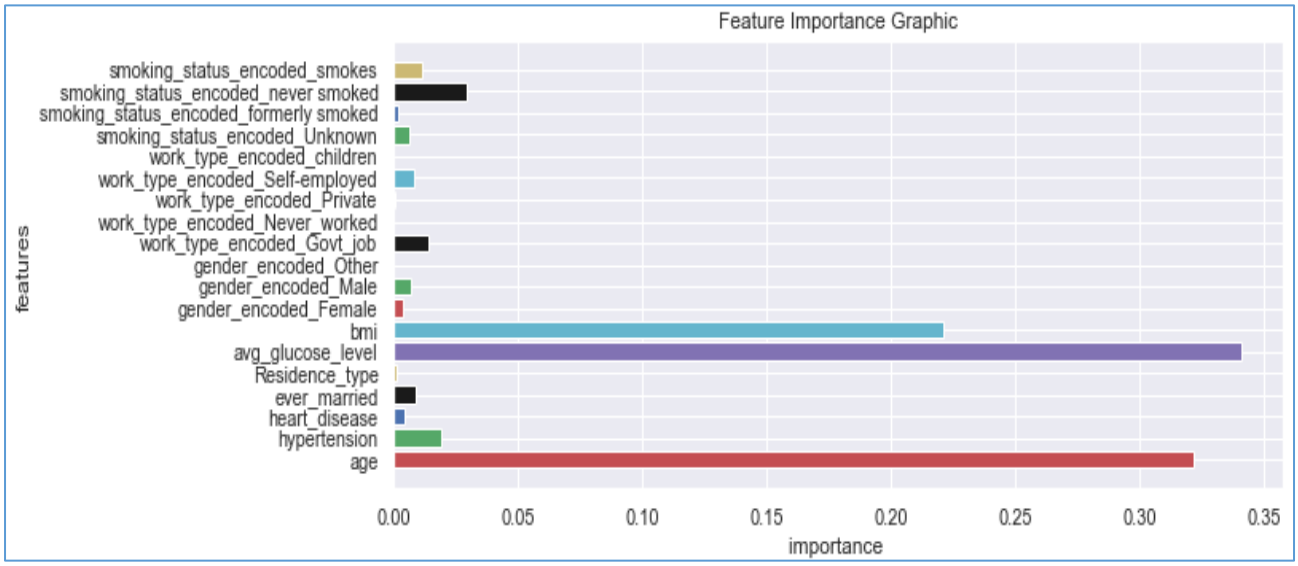

Figure 8; Shows Feature Importance Computed by HFE Algorithm

As presented in Figure 8, the importance of features if provided in horizontal axis and vertical axis shows all the identified features from the given dataset [21]. There are 19 features identified by the algorithm and out of which 7 features are found to have significant contribution towards brain stroke detection.

4.3. Performance Evaluation

This section presents the performance evaluation with all the brain stroke prediction models in terms of precision, recall, F1-measure and accuracy.

Table 4. Shows Performance of Brain Stroke Prediction Models

Brain Stroke Prediction Model	Performance (%)			
	Precision	Recall	F1-Measure	Accuracy
GaussianNB	0.88	0.18	0.1	0.18
BernoulliNB	1	0.97	0.98	0.97
Logistic Regression	1	0.97	0.98	0.97
RandomForest Classifier	1	0.97	0.98	0.97
Support Vector Machine	1	0.97	0.98	0.97
Decision Tree Classifier	0.93	0.94	0.93	0.94
KNeighbors Classifier	1	0.97	0.98	0.97
Gradient Boosting Classifier	1	0.97	0.98	0.97
Stochastic Gradient Descent	1	0.97	0.98	0.97
Neural Nets	0.97	0.96	0.96	0.96

As presented in Table 4, the prediction models and their performance when HEF is used is provided in terms of precision, recall, F1-measure and accuracy.

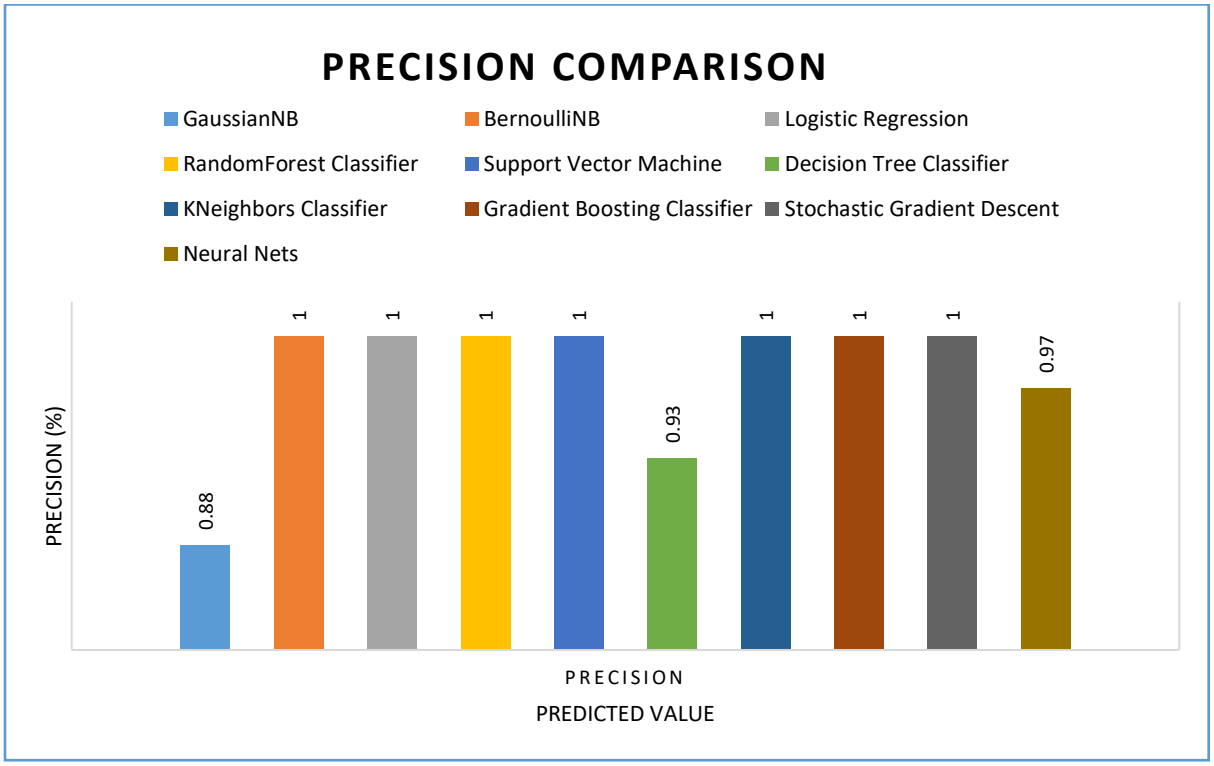

Figure 9; Performance Comparison in Terms of Precision

As presented in Figure 9, the precision performance of various brain stroke prediction models is provided. Each model has its modus operandi and thus has its performance in terms of precision. Interestingly many prediction models showed highest precision that is 1. The least performance is exhibited by GaussianNB with 0.88. The precision of DT is 0.93 and Neural Nets is 0.97.

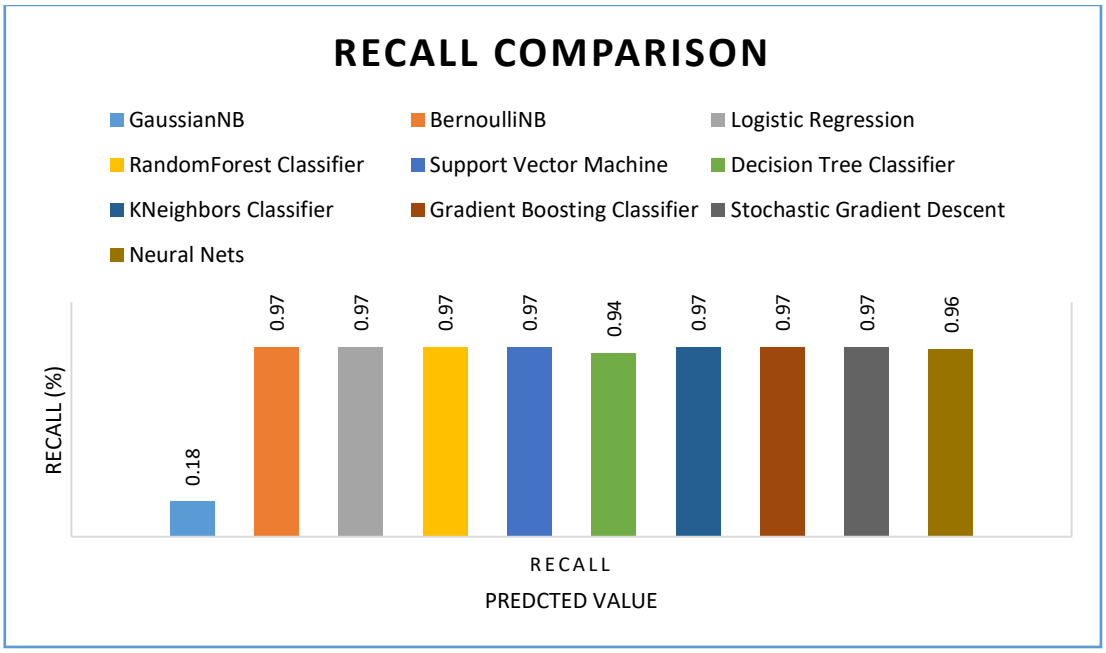

Figure 10; Performance Comparison in Terms of Recall

As presented in Figure 10, the recall performance of various brain stroke prediction models is provided. Each model has its modus operandi and thus has its performance in terms of recall. Interestingly many prediction models showed highest recall that is 0.97. The least performance is exhibited by GaussianNB with 0.18. The recall of DT is 0.94 and Neural Nets is 0.96.

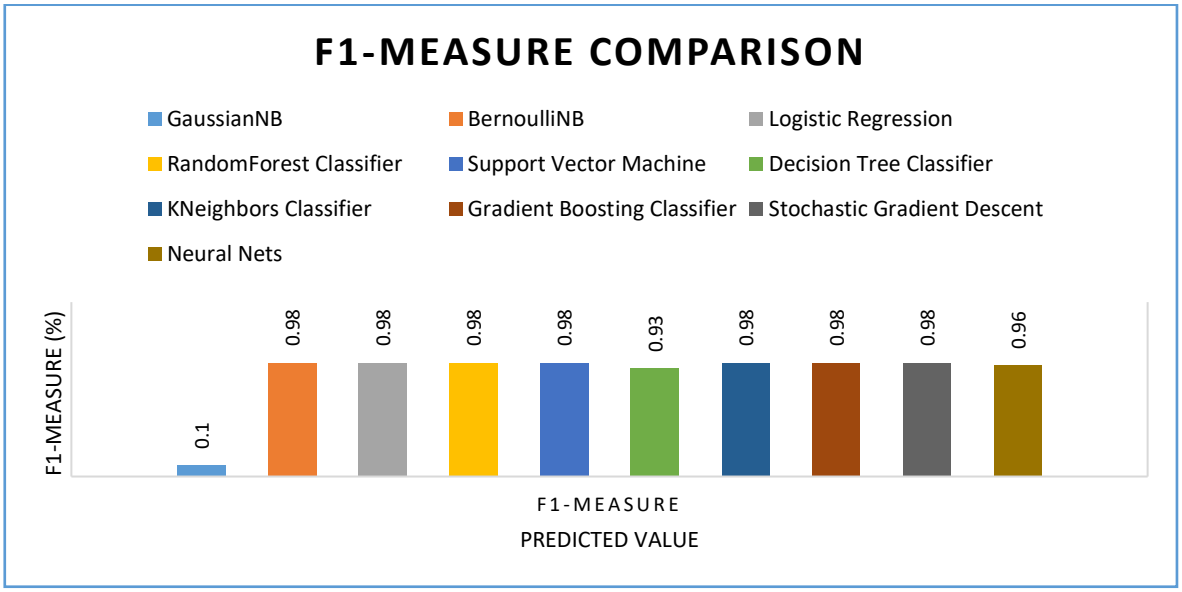

Figure 11; Performance Comparison in Terms of F1-measure

As presented in Figure 11, the F1-measure performance of various brain stroke prediction models is provided. Each model has its modus operandi and thus has its performance in terms of F1-measure. Interestingly many prediction models showed highest F1-measure that is 0.98. The least performance is exhibited by GaussianNB with 0.1. The F1-measure of DT is 0.93 and Neural Nets is 0.96.

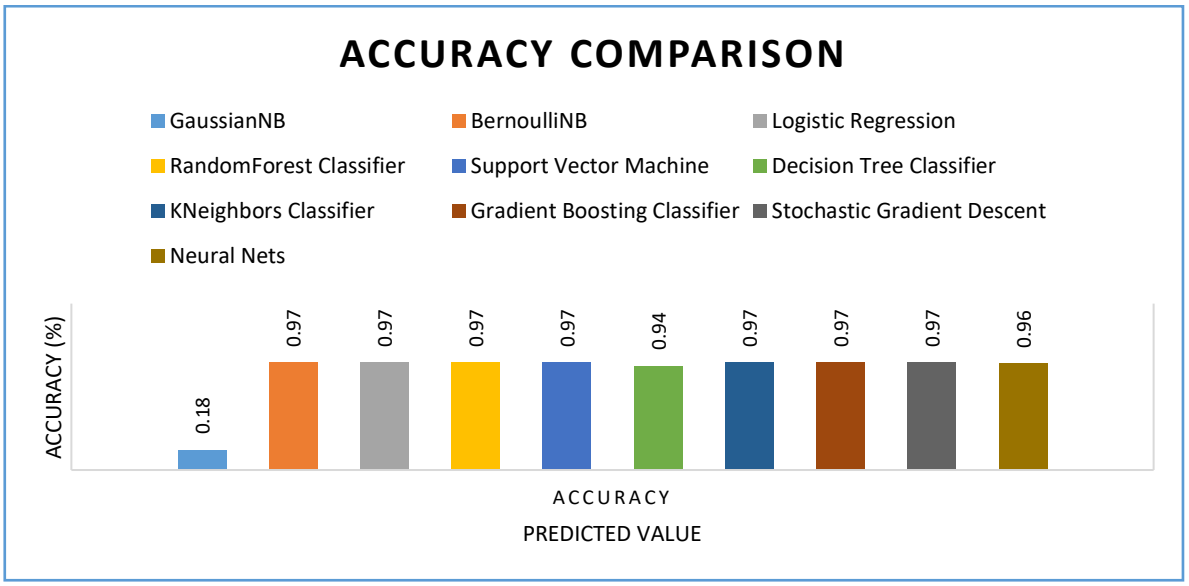

Figure 12; Performance Comparison in Terms of Accuracy

As presented in Figure 12, the accuracy performance of various brain stroke prediction models is provided. Each model has its modus operandi and thus has its performance in terms of accuracy. Interestingly many prediction models showed highest accuracy that is 0.97. The least performance is exhibited by GaussianNB with 0.18. The accuracy of DT is 0.94 and Neural Nets is 0.96.

Table 5. Performance Comparison of Prediction Models with and without Feature Engineering

Brain Stroke Predicted Model	Accuracy (%)	
	Without HFE	With HFE
GaussianNB	0.1746	0.18
BernoulliNB	0.9409	0.97
Logistic Regression	0.9409	0.97
RandomForest Classifier	0.9409	0.97
Support Vector Machine	0.9409	0.97
Decision Tree Classifier	0.9118	0.94
KNeighbors Classifier	0.9409	0.97
Gradient Boosting Classifier	0.9409	0.97
Stochastic Gradient Descent	0.9409	0.97
Neural Nets	0.9312	0.96

As presented in Table 5, the brain stroke prediction models are provided with their accuracy performance with and without feature engineering.

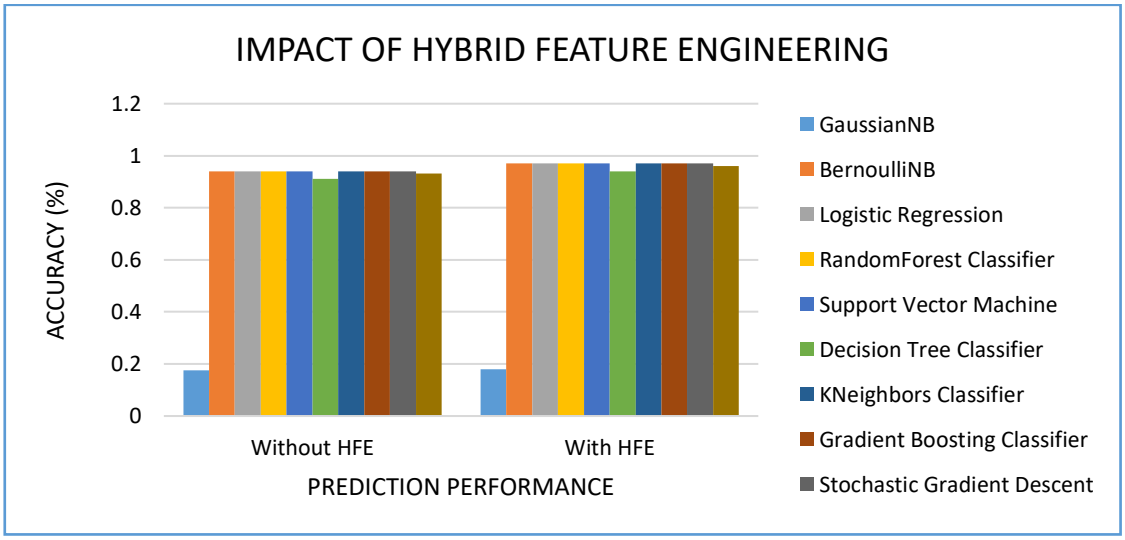

Figure 13; Performance Evaluation of Brain Stroke Prediction Models in Terms of Accuracy

As presented in Figure 13, the performance of many brain stroke prediction models made up of supervised learning technique is evaluated. The observations are made in terms of accuracy (%) when the proposed HFE is used and without the HFE. It is evident from the performance

of the models that HFE has its significant impact on the prediction models. The least performing model is GaussianNB. Other prediction models do have comparable and acceptable performance. Without the proposed HFE algorithm, BernoulliNB showed 94.09% accuracy, Logistic Regression 94.09%, Random Forest 94.09%, SVM 94.09%, DT91.18%, KNeighbors 94.09%, Gradient Boosting 94.09%, SGD 94.09% and Neural Nets 93.12%. With feature engineering, the same prediction models showed improved accuracy. With the feature engineering BernoulliNB showed 97.0% accuracy, Logistic Regression 97.0%, Random Forest 97.0%, SVM 97.0%, DT 94.0%, KNeighbors 97.0%, Gradient Boosting 97.0%, SGD 97.0% and Neural Nets 96.0%. Average improvement in accuracy of the models with the proposed HFE is 2.896666667% which is significant considering the sensitivity associated with the medical data.

5. Conclusion and Future Work

In this paper, we proposed a ML framework known as Brain Stroke Detection Framework (BSDF). We also proposed a hybrid feature engineering method that will be used in ML pipeline of the framework for leveraging prediction performance. The algorithm is known as Hybrid Feature Engineering (HFE) which is the combination of three filter based approaches. The framework is realized with another algorithm known as Supervised Machine Learning Models for Brain Stroke Detection (SML-BSD) which exploits HFE for improving prediction performance. It is a data driven approach to have cheaper alternative to complement Clinical Decision Support System (CDSS) in healthcare units. A prototype application is built using Python data science platform to evaluate the proposed framework and underlying algorithms with benchmark dataset and the dataset collected from a corporate healthcare unit. Many brain stroke prediction models could achieve 97% accuracy when HFE is used as underlying feature selection method. There is significant improvement in performance of different brain stroke prediction models with the hybrid feature engineering algorithm. In future there are several possibilities to improve the research on brain stroke detection. Two important directions are provided here. First, along with feature engineering ensemble mechanisms can be exploited to improve prediction performance further. Second, there is need for exploiting deep learning alternatives with brain MRI scans towards brain stroke detection.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1]. Manisha Sanjay Sirsat, Eduardo Ferm e and Joana C amara. (2020). Machine Learning for Brain Stroke: A Review. Stroke and Cerebrovascular Diseases. 29, p1-17.

[2]. Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, Md. Mahfujur Rahman, M Shamim Al Mamun and M Shamim Kaiser. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. IEEE, p1-6.

[3]. S. Pradeepa, K.R. Manjula, S Vimal, Mohammad S, Khan, Naveen Chilamkurti and Ashish Kr. Luhach. (2020). DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. Springer Science, p1-19.

[4]. Pawan Kumar Tiwari, Mukesh Kumar Yadav, R. K. G. A. . (2022). Design Simulation and Review of Solar PV Power Forecasting Using Computing Techniques. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(5), 18–27. https://doi.org/10.17762/ijrmee.v9i5.370

[5]. Haris Kamal, Victor Lopez and Sunil A. Sheth. (2018). Machine Learning in Acute Ischemic Stroke Neuroimaging. Frontiers in Neurology, p1-6.

[6]. Osaleena Mohanty, Anita M. Sinha, Alexander B. Remsik , Keith C. Dodd , Brittany M. Young, Tyler Jacobson, Matthew McMillan, Jaclyn Thoma, Hemali Advani and Veena A. Nair. (2018). Machine Learning Classification to Identify the Stage of Brain-Computer Interface Therapy for Stroke Rehabilitation Using Functional Connectivity. Frontiers in Neurology, p1-14.

[7]. Kadhim, R. R., and M. Y. Kamil. "Evaluation of Machine Learning Models for Breast Cancer Diagnosis Via Histogram of Oriented Gradients Method and Histopathology Images". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 36-42, doi:10.17762/ijritcc.v10i4.5532.

[8]. Priya Govindarajan, Ravichandran Kattur Soundarapandian, Amir H. Gandomi, Rizwan Patan, Premaladha Jayaraman and Ramachandran Manikandan. (2019). Classification of stroke disease using machine learning algorithms. Intelligent Biomedical Data Analysis and Processing, 1-12.

[9]. Marco Salucci, Jan Vrba, IljaMerunka and Andrea Massa. (2017). Real-time brain stroke detection through a learning-by-examples technique An experimental assessment. DIGITEO, p2796-2799.

[10]. Hyunna Lee, PhD Eun-Jae Lee, MD Sungwon Ham, MS Han-Bin Lee, MD Ji Sung Lee, Sun U. Kwon, Jong S. Kim, Namkug Kim and PhD Dong-Wha Kang MD. (2020). Machine Learning Approach to Identify Stroke Within 4.5 Hours. Original Contribution, p1-7.

[11]. Yoon-A Choi, Sejin Park, Jong-Arm Jun, Chee Meng Benjamin Ho, Cheol-Sig Pyo, Hansung Lee and Jaehak Yu. (2021). Machine-Learning-Based Elderly Stroke Monitoring System Using Electroencephalography Vital Signals. Applied Sciences, p1-18.

[12]. Tessy Badriyah, Nur Sakinah and Iwan Syarif. (2020). Machine Learning Algorithm for Stroke Disease Classification. Electrical, Communication and Computer Engineering, p1-5.

[13]. Md. Monirul Islam, Sharmin Akter and Md. Rokunojjaman. (2021). Stroke prediction analysis using machine learning classifiers and feature technique. Electronics and Communications System. 1, p1-7.

[14]. Gang Fanga, Wenbin Liu and Lixin Wang. (2020). A machine learning approach to select features important to stroke prognosis. Computational Biology and Chemistry, p1-9.

[15]. Gupta, D. J. . (2022). A Study on Various Cloud Computing Technologies, Implementation Process, Categories and Application Use in Organisation. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 09–12. https://doi.org/10.17762/ijfrcsce.v8i1.2064

[16]. Suganya, TS and Murugavalli, S (2017). International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET) -

Feature selection for an automated ancient Tamil script classification system using machine learning techniques, p1–6.

[17]. Açikoglu, Merve and Tuncer Seda Arslan (2019). Incorporating Feature Selection Methods into a Machine Learning-Based Neonatal Seizure Diagnosis. Medical Hypotheses, p1-21.

[18]. Chen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee. (2017). Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database. IEEE, p3110-3113.

[19]. M. . Parhi, A. . Roul, B. Ghosh, and A. Pati, "IOATS: an Intelligent Online Attendance Tracking System based on Facial Recognition and Edge Computing", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 252–259, May 2022.

[20]. Ray, Sujan; Alshouiliy, Khaldoon; Roy, Anupam; AlGhamdi, Ali and Agrawal, Dharma P. (2020). 2020 Intermountain Engineering, Technology and Computing (IETC) - Chi-Squared Based Feature Selection for Stroke Prediction using AzureML, p1–6.

[21]. Salucci, Marco; Marcantonio, Davide; Li, Maokun; Oliveri, Giacomo; Rocca, Paolo and Massa, Andrea (2019). 2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS) - Innovative Machine Learning Techniques for Biomedical Imaging, p1–3.

[22]. Suresh, Shruthi and Duerstock, Bradley S. (2018). IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) - Optimal Feature Selection for the Detection of Autonomic Dysreflexia in Individuals with Tetraplegia, p480–485.

[23]. Kai Keng Ang, Zhang Yang Chin, Haihong Zhang and Cuntai Guan, (2008). IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) - Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface, (0), p2390–2397.

[24]. Sung, Sheng-Feng; Lin, Chia-Yi and Hu, Ya-Han (2020). EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. IEEE Journal of Biomedical and Health Informatics, 1-12.

[25]. Brain Stroke Dataset, Retrieved from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[26]. Brain Stroke Dataset, collected from https://mallareddynarayana.com/

[27]. Maldorad S, Weber R (2009) A wrapper method for feature selection using support vector. Machines Information Sciences 179:2208–2217.

[28]. Rice JA (2006). Mathematical Statistics and Data Analysis. Third Edition.

[29]. Kullback, S.; Leibler, R.A. (1951). "On Information and Sufficiency". Annals of Mathematical Statistics 22 (1): 79–86.

[30]. Garg, K. . (2022). Compactness in General Category Theory. International Journal on Recent Trends in Life Science and Mathematics, 9(1), 19–27. https://doi.org/10.17762/ijlsm.v9i1.138