Healthcare Analytics 2 (2022) 100032

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A predictive analytics approach for stroke prediction using machine learning and neural networks

Soumyabrata Dev [a,b,*], Hewei Wang [c,d], Chidozie Shamrock Nwosu [e], Nishtha Jain [a], Bharadwaj Veeravalli [f], Deepu John [g]

[a] ADAPT SFI Research Centre, Dublin, Ireland
[b] School of Computer Science, University College Dublin, Ireland
[c] Beijing University of Technology, Beijing, China
[d] Beijing-Dublin International College, Beijing, China
[e] National College of Ireland, Dublin, Ireland
[f] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[g] School of Electrical and Electronic Engineering, University College Dublin, Ireland

ARTICLE INFO

Keywords:
predictive analytics
machine learning
neural network
electronic health records
stroke

ABSTRACT

The negative impact of stroke in society has led to concerted efforts to improve the management and diagnosis of stroke. With an increased synergy between technology and medical diagnosis, caregivers create opportunities for better patient management by systematically mining and archiving the patients' medical records. Therefore, it is vital to study the interdependency of these risk factors in patients' health records and understand their relative contribution to stroke prediction. This paper systematically analyzes the various factors in electronic health records for effective stroke prediction. Using various statistical techniques and principal component analysis, we identify the most important factors for stroke prediction. We conclude that age, heart disease, average glucose level, and hypertension are the most important factors for detecting stroke in patients. Furthermore, a perceptron neural network using these four attributes provides the highest accuracy rate and lowest miss rate compared to using all available input features and other benchmarking algorithms. As the dataset is highly imbalanced concerning the occurrence of stroke, we report our results on a balanced dataset created via sub-sampling techniques.

1. Introduction

We have witnessed amazing developments in the field of medicine with the aid of technology [1]. With the advent of annotated dataset of medical records, we can now use data mining techniques to identify trends in the dataset. Such analysis has helped the medical practitioners to make an accurate prognosis of any medical conditions. It has led to an improved healthcare conditions and reduced treatment costs. The use of data mining techniques in medical records have great impact on the fields of healthcare and bio-medicine [2,3]. This assists the medical practitioners to identify the onset of disease at an earlier stage. We are particularly interested in stroke, and to identify the key factors that are associated with its occurrence.

Several studies [4–7] have analysed the importance of lifestyle types, medical records of patients on the probability of the patients to develop stroke. Further, machine learning models are also now employed to predict the occurrence of stroke [8,9]. However, there is no study that attempts to analyse all the conditions related to patient, and identify the key factors necessary for stroke prediction. In this paper, we attempt to bridge this gap by providing a systematic analysis of the various patient records for the purpose of stroke prediction. Using a publicly available dataset of 29072 patients' records, we identify the key factors that are necessary for stroke prediction. We use principal component analysis (PCA) to transform the higher dimensional feature space into a lower dimension subspace, and understand the relative importance of each input attributes. We also benchmark several popular machine-learning based classification algorithms on the dataset of patient records.

The main contributions of this paper are as follows – (a) we provide a detailed understanding of the various risk factors for stroke prediction. We analyse the various factors present in Electronic Health Record (EHR) records of patients, and identify the most important factors necessary for stroke prediction; (b) we also use dimensionality

* Corresponding author at: School of Computer Science, University College Dublin, Ireland.
E-mail address: soumyabrata.dev@ucd.ie (S. Dev).

[1] In the spirit of reproducible research, the code and data to reproduce the results in this manuscript are available online here: https://github.com/Soumyabrata/EHR-features.

https://doi.org/10.1016/j.health.2022.100032
Received 28 December 2021; Received in revised form 4 February 2022; Accepted 9 February 2022

reduction technique to identify patterns in low-dimension subspace of the feature space; and (c) we benchmark popular machine learning models for stroke prediction in a publicly available dataset. We follow the spirit of reproducible research, and therefore the source code of all simulations used in this paper are available online.[1]

The structure of the paper is as follows. The remaining part of Section 1 provides an overview of the related work, and describes the dataset used in our study. Section 2 covers the correlation analysis and feature importance analysis. The results from Principal Component Analysis are explained in Section 3. The data mining algorithms used for predictive modelling and their performance on the dataset is detailed in Section 4. Finally, Section 5 concludes the paper and discusses future work.

1.1. Related work

Existing works in the literature have investigated various aspects of stroke prediction. Jeena et al. provides a study of various risk factors to understand the probability of stroke [8]. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. In Hanifa and Raja [9], an improved accuracy for predicting stroke risk was achieved using radial basis function and polynomial functions applied in a non-linear support vector classification model. The risk factors identified in this work were divided into four groups — demographic, lifestyle, medical/clinical and functional. Similarly, Luk et al. studied 878 Chinese subjects to understand if age has an impact on stroke rehabilitation outcomes [10]. Min et al. in [11] developed an algorithm for predicting stroke from potentially modifiable risk factors. Singh and Choudhary in [12] have used decision tree algorithm on Cardiovascular Health Study (CHS) dataset for predicting stroke in patients. A deep learning model based on a feed-forward multi-layer artificial neural network was also studied in [13] to predict stroke. Similar work was explored in [14–16] for building an intelligent system to predict stroke from patient records. Hung et al. in [17] compared deep learning models and machine learning models for stroke prediction from electronic medical claims database. In addition to conventional stroke prediction, Li et al. in [18] used machine learning approaches for predicting ischaemic stroke and thromboembolism in atrial fibrillation.

The results from the various techniques are indicative of the fact that multiple factors can affect the results of any conducted study. These various factors include the way the data was collected, the selected features, the approach used in cleaning the data, imputation of missing values, randomness and standardization of the data will have an impact on the outcome of any study carried. Therefore, it is important for the researchers to identify how the different input factors in an electronic health record are related to each other, and how they impact the final stroke prediction accuracy.

Studies in related areas [3,19] demonstrate that identifying the important features impacts the final performance of machine learning framework. It is important for us to identify the perfect combination of features, instead of using all the available features in the feature space. As indicated in [3], redundant attributes and/or totally irrelevant attributes to a class should be identified and removed before the use of a classification algorithm. Therefore, it is essential for data mining practitioners in healthcare to identify how the risk factors captured in electronic health records are inter-dependent, and how they impact the accuracy of stroke prediction independently.

1.2. Electronic health records dataset

An Electronic Health Record (EHR) also known as Electronic Medical Record (EMR), is a repository of information for a patient. It is an automated, computer readable storage of the medical status of a patient that is keyed in by qualified medical practitioners. The records contain vitals, diagnosis or medical exam results of a patient. The future of medical diagnosis looks promising with the optimal use of EHR. The use of EHR increased from 12.5% to 75.5% in US Hospitals between 2009 and 2014 as indicated by the statistics recorded in [20].

For our study, we use a dataset of electronic health records released by McKinsey & Company as a part of their healthcare hackathon challenge.[2] The dataset is available from Kaggle,[3] a public data repository for datasets. The dataset contains the EHR records of 29072 patients. It has a total of 11 input attributes, and 1 output feature. The output response is a binary state, indicating if the patient has suffered a stroke or not. The remaining 11 input features in EHR are: patient identifier, gender (G), age (A), binary status if the patient is suffering from hypertension (HT) or not, binary status if the patient is suffering from heart disease (HD) or not, marital status (M), occupation type (W), residence (urban/rural) type (RT), average glucose level (AG), body mass index (BMI), and patient's smoking status (SS). The dataset is highly unbalanced with respect to the occurrence of stroke events; most of the records in the EHR dataset belong to cases that have not suffered from stroke. The publisher of the dataset has ensured that the ethical requirements related to this data are ensured to the highest standards. In the subsequent discussion of this paper, we will exclude the patient identifier as one of the input feature. We will consider the remaining 10 input features, and 1 response variable, in our study and analysis.

2. Analysing electronic health records

In this section, we provide an analysis of electronic health records dataset. We perform correlation analysis of the features. We use the entire dataset of EHR records to perform such analysis on the input features of the EHR records. Correlation analysis is useful for feature selection in the following manner: if two features have very high correlation, one of them can be ignored in the prediction of occurrence of stroke as it does not contribute any additional knowledge to the prediction model. Moreover, we evaluate the behaviour of the features individually and in a group to gaze the importance of each individual feature in predicting the occurrence of a stroke. A systematic analysis of the input feature space is an integral part for stroke prediction. It is important to find the optimal and minimal set of predictive features to reduce the computational cost of modelling and efficient archival of EHR records. This paves us the path for clinicians to record *only* those features in the EHR records that are most efficient for stroke prediction.

2.1. Correlation between features

We use Pearson's correlation coefficient to generate Fig. 1, which shows the correlation between different patient attributes. The strength of the linear relationship between any two features of the patient's electronic health data will be determined by this correlation value. We have used a colourmap in Fig. 1, such that the blue colour represents positive correlation, while red is negative. The deeper the colour and larger the circle size, the higher is the correlation between the two patient attributes.

As is intuitive, the correlation of an attribute with itself is unity. There is a significant correlation between a patient's marital status and their age with 0.5 correlation index. There is also a positive correlation between patient's age and the type of their work with 0.38 correlation index, whether they suffer from hypertension and heart disease or not and their average glucose level. This correlation of patient's age with other attributes seems intuitive, as most ailments occur in an ageing population. The type of residence of patient is not correlated with any other attribute. Patient's type of work has a positive correlation with their marital status with 0.35 correlation index.

[2] https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/.

[3] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

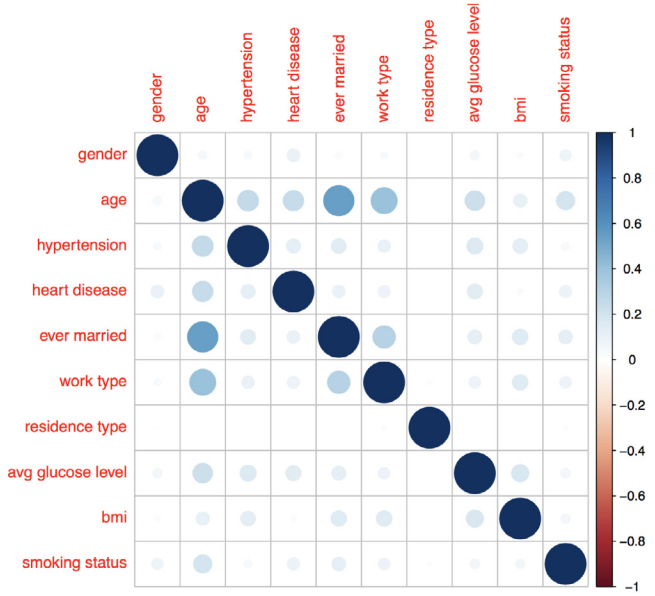


Fig. 1. Correlation matrix for patient attributes in the dataset. These attributes are gender, age, status 0/1 if patient is suffering from hypertension, status 0/1 if patient is suffering from heart disease, marital status, work type, residence type, average glucose level, body mass index and patient's smoking status.

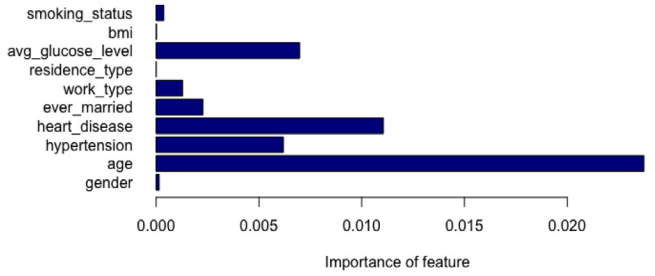


Fig. 2. Importance of patient attributes in predicting the occurrence of stroke with a Linear Vector Quantization (LVQ) model.

In summary, the correlation matrix shown in Fig. 1 tells us that none of the features are highly correlated with each other. Thus, each feature might have their individual contribution towards stroke prediction. The next two subsections analyse the importance of an individual feature for stroke prediction.

2.2. Individual features for stroke prediction

Fig. 2 shows the importance of each patient's attribute in predicting the occurrence of stroke using a Learning Vector Quantization (LVQ) model. The relative importance of a patient's attribute is measured by the increase in the model's prediction error due to that attribute. We use the varImp method from the R `caret` package[4] to compute this relative feature importance. As Fig. 2 illustrates, patient's age (A) is the feature with highest importance in predicting the occurrence of stroke. The other features with high importance are presence of heart disease (HD), patient's average glucose level (AG) and presence of hypertension.

The analysis described above shows patient's age (A) has a comparatively higher importance by itself, yet a combination of different features may improve prediction because they are not correlated with each other. Furthermore, we also compute the $CHADS_2$ score for the EHR records. CHADS2 score is a stroke risk score for non valvular atrial

[4] http://topepo.github.io/caret/index.html.

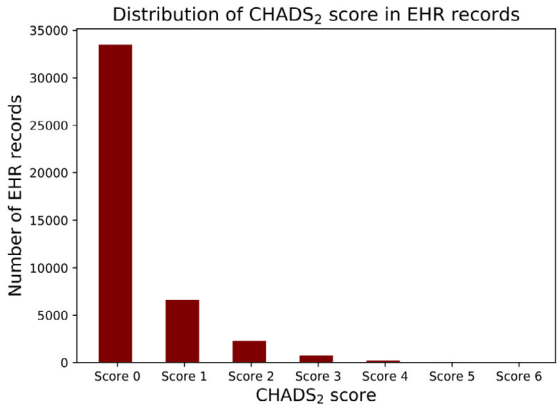


Fig. 3. We show the distribution of $CHADS_2$ score in the EHR records dataset. We observe that most of the $CHADS_2$ score values are low.

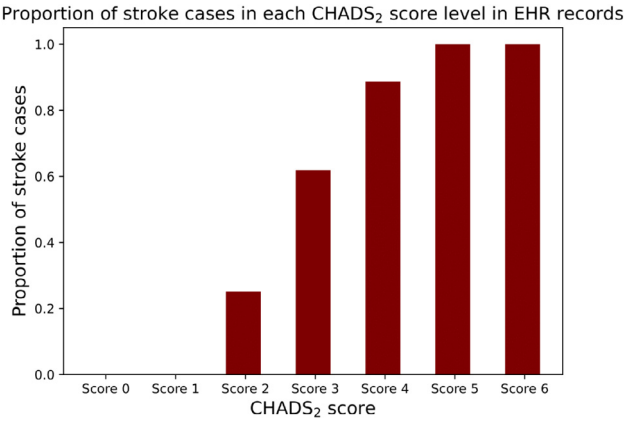


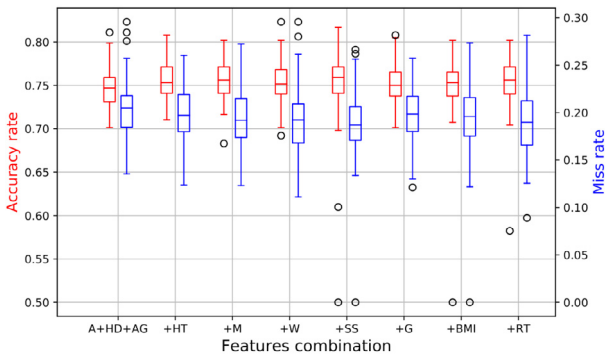
Fig. 4. The proportion of cases with a stroke event as predicted by $CHADS_2$ for each score level respectively.

fibrillation, where C represents congestive heart failure or impairment of left ventricular function, H represents whether the patient has hypertension, A stands for age, D stands for diabetes, S stands for stroke, or transient ischaemic attack, history of thromboembolism. Fig. 3 shows the distribution of the $CHADS_2$ score for our dataset. We observe that most of the EHR observations have a low $CHADS_2$ score. Fig. 4 shows the proportion of cases with a stroke event as predicted by CHADS2 for each score level. We observe that the larger the CHADS2 score, the higher is the occurrence of stroke cases. Combined with Figs. 3 and 4, we find that most people having $CHADS_2$ score of 1 or 2, have a low probability of stroke. We observe that only a small number of people with $CHADS_2$ score greater than 2 have a higher probability of occurrence of stroke.

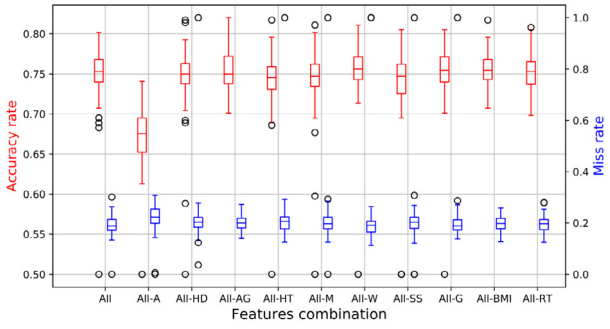
2.3. Selection of optimum features for stroke prediction

In the previous section, we used the features one at a time and saw that there are only few features which have higher importance in stroke prediction. In this section, we use all features then subsequently remove one feature at a time or add one feature at a time and further analyse the importance of an individual feature in stroke prediction. We use neural network algorithm to produce the results. We use a perceptron neural network for this experiment. We use the entire dataset of EHR records to perform the feature analysis.

Fig. 5(a) shows the results for subsequently adding one feature at a time. The first result corresponds to using features A, HD and AG. These features are chosen as from Fig. 2, we observe that these three features show higher importance compared to others. When



(a) Adding features



(b) Removing features

Fig. 5. We measure the accuracy rate and miss rate of stroke prediction when (a) when features are added one at a time, and (b) individual features are deleted one at a time. The box plots represent the distribution of the metrics, computed from 100 experiments.

HT is added to this pool, we see that the accuracy rate is slightly improved and miss rate is slightly decreased. Note that HT is the fourth important variable from the analysis of Fig. 2. Subsequently, when other features are added one by one, the accuracy rate and miss rate show very slight variation. The accuracy and miss rate are almost same for all the remaining configurations. Therefore, it suggests that the four features: A, HD, AG and HT can be optimum features as there is no improvement offered by addition of other features.

Fig. 5(b) shows the results for deleting one feature at a time from the pool of all features. Here we can observe that the accuracy rate is significantly affected when the feature A is removed from the pool. This is inline with our earlier discussion which showed that A has highest score amongst all features (ref to Fig. 2). There is no much significant changes when other features are removed from the pool.

3. Principal component analysis

In this section, we analyse the variance in the dataset using Principal Component Analysis (PCA). In this multivariate analysis, the dataset is transformed into a set of values of linearly uncorrelated variables called principal components such that maximum variance is extracted from the variables. These principal components act as summaries of the features of the dataset. These new basis functions do not have a physical interpretation. However, these new basis functions are linear combinations of the original feature vectors. In this work, we do not restrict ourselves on the feature analysis using traditional feature elimination techniques. However, we use dimensionality reduction technique to transform the high-dimensional feature space onto 2-dimensional feature space to understand the inter-relation amongst the feature space. PCA can be used to reduce the feature space for predictive modelling if the first few components capture most of the variance in the data. We analyse the 10 dimensional patient attribute in the lower dimension subspace using PCA. We provide a brief primer on principal component analysis and mathematically formulate our problem statement.

Let us suppose that $\mathbf{X}$ is the variable matrix of dimension $m \times n$. In this case, m indicates the total number of input attributes in EHR, and n is the total number of patient records in the dataset. Therefore, $m = 10$ and $n = 29072$ in this analysis for stroke prediction. We vectorize the individual features f_{1-10} from the matrix $\mathbf{X}$, into corresponding $\widetilde{\mathbf{v}}_j \in \mathbb{R}^{mn\times 1}$ where $j = 1, 2, \ldots, 10$. Finally, the $\widetilde{\mathbf{v}}_j$ features are stacked together to create the matrix $\hat{\mathbf{X}} \in \mathbb{R}^{mn\times 10}$:

$$\hat{\mathbf{X}} = [\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \widetilde{\mathbf{v}}_3, \ldots, \widetilde{\mathbf{v}}_{10}]. \tag{1}$$

We perform the PCA on the normalized matrix of $\hat{\mathbf{X}}$, that is normalized using the corresponding means $\bar{v}_j$ and standard deviations σ_{v_j} of the individual features. The normalized matrix $\ddot{\mathbf{X}}$ is represented as:

$$\ddot{\mathbf{X}} = \left[\frac{\widetilde{\mathbf{v}}_1 - \bar{v}_1}{\sigma_{v_1}}, \frac{\widetilde{\mathbf{v}}_2 - \bar{v}_2}{\sigma_{v_2}}, \ldots, \frac{\widetilde{\mathbf{v}}_j - \bar{v}_j}{\sigma_{v_j}}, \ldots, \frac{\widetilde{\mathbf{v}_{10}} - \bar{v_{10}}}{\sigma_{v_{10}}}\right]. \tag{2}$$

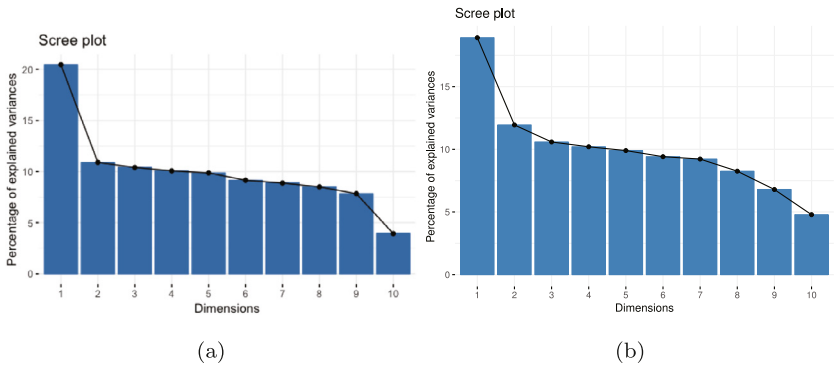
We interpret how the results from PCA are related to predictive variables or features represented by patient attributes and the individual observations represented by the medical health records. We study the relation between the first two principal components and the individual input variables. We also study the importance of the first two principal components for a given observation. We use the guide by Abdi and Williams [21] for this study.

3.1. Variance explained by principal components

A scree plot is used to select the components which explain most variability in the dataset, generally 80% or more variance. Fig. 6(a) shows the percentage of variance in the dataset explained by the different principal components in the original dataset. Here, we can observe that the variance explained by different principal components are very low. Out of 10 principal components, 8 are needed to explain variance of 88.2%. Balanced dataset means that the dataset is balanced with respect to the stroke labels using random sampling. The original dataset is unbalanced because there are more samples possessing negative stroke labels, as compared to positive label stroke samples. We make it balanced by considering all the positive stroke samples, and then randomly picking equal number of negative stroke samples from the rest. This will make a balanced dataset with equal number of positive and negative stroke samples.

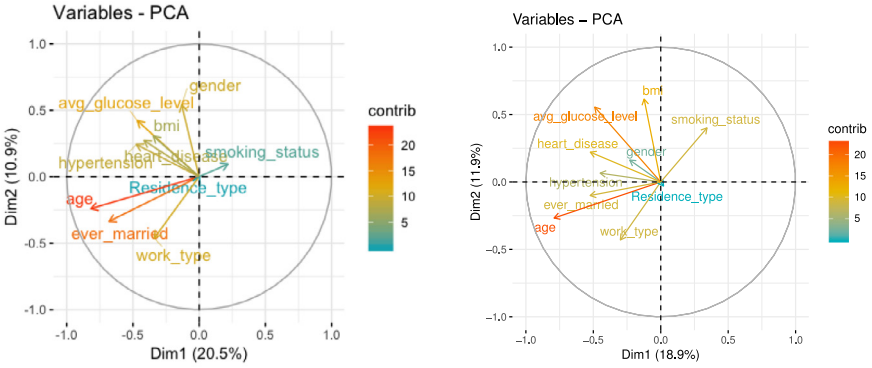
All principal components are orthogonal to each other and hence uncorrelated. Therefore, each individual PC can be useful to explain a unique phenomenon. The distributed variance in Fig. 6 indicates that the different principal components are explaining different underlying phenomenon. These phenomenon can be analysed based on the variable loadings. Variable loadings are the contribution of different variables to an individual principal component. We have included the scree plot for the balanced dataset as well in Fig. 6(b). This is useful for researchers to understand the impact of unbalanced nature of the dataset on the subspace representation. Table 1 shows the contribution of each variable towards first two principal components. The sum of the squares of all loadings for an individual principal component equals unity. Therefore, if the variable loading crosses a threshold value of $\sqrt{1/10} = 0.31$, it indicates that the variable has a strong contribution towards the principal component. In Table 1, the variables that cross the threshold are in bold. Here we observe that variables like A, AG, HT and M have strong contribution towards the first principal component. The variable HD also shows significant contribution towards it. From earlier discussions, we saw that these features actually are important from stroke prediction point of view. Therefore, it indicates that the first principal component (which has stronger loadings from these variables) might be useful in predicting the stroke.

In the following sections, we will assess the role of these principal components in stroke prediction.



(a) (b)

Fig. 6. Percentage of variance explained by different principal components. We show the scree plot for (a) original, (b) balanced datasets.



(a) (b)

Fig. 7. Biplot representation of the input attributes on the first two principal components. We show the biplot for (a) original, (b) balanced datasets.

Table 1
We check the contributions of the different features in the first and second principal components. We report the absolute values of the different loading factors.

Features	PC_1	PC_2
gender	0.092	**0.516**
age	**0.571**	0.230
hypertension	**0.331**	0.232
heart_disease	0.290	0.261
ever_married	**0.475**	**0.322**
work_type	0.236	**0.442**
residence_type	0.001	0.003
avg_glucose_level	**0.326**	**0.403**
bmi	0.242	0.293
smoking_status	0.152	0.090

3.2. Relation between principal components with patient attributes

Fig. 7 describes the biplot representation. It shows how the different input attributes are correlated with other, and also depend on the first and second principal components. The x-axis and y-axis indicate the first and second principal components respectively. The each vectors represent an input attribute, and its length indicate its importance. We observe that the average glucose level and the heart disease are correlated to each other. The age has the biggest contribution in the first two principal component. We also observe that the orientation of the different feature vectors in the two-dimensional feature space is the same as the unbalanced dataset.

Fig. 7(a) illustrates that the contribution of patient's residence type is minimum to the two principal components. We also observe that age and status of marriage are correlated with each other, and have a high contribution to the first principal component. The smoking status of a patient and its average glucose level are orthogonal to each other, indicating that they provide different information to the feature space. However, smoking status and age point opposite to each other, indicating that they provide similar information, but have a diverging characteristics. We also compute the biplot for the EHR records on the balanced dataset. We show the corresponding biplot in Fig. 7(b). We observe that the average glucose level and the heart disease are correlated to each other. The age has the biggest contribution in the first two principal component. We also observe that the orientation of the different feature vectors in the two-dimensional feature space is the same as the unbalanced dataset.

3.3. Relation of principal components with individual records

Finally, we also check the relation of individual patient record observations on the first two principal components. In Fig. 8, each dot represents a patient record observation.

Fig. 8(a) shows the importance of principal components onto each of the records. This is generally represented by the *cos2* measure, indicating the squared distance from the origin. This indicates that observations that possess a high *cos2* value can be represented by the principal components, as opposed to observations with a low *cos2* value. We observe that a significant number of points are clustered near the origin, which cannot be represented completely by the principal components. Therefore, the principal component can indicate only a part of the entire information in the feature space.

Fig. 8(b) colour codes the patient records based on the status of the stroke. As the dataset is highly unbalanced, we observe that most of the observations are colour coded with negative status of stroke. We observe a few observations with positive status of stroke (observations with 1 label). However, these positive stroke observations are not located in clusters. This indicates that higher-dimensional features are necessary to separate the observations.

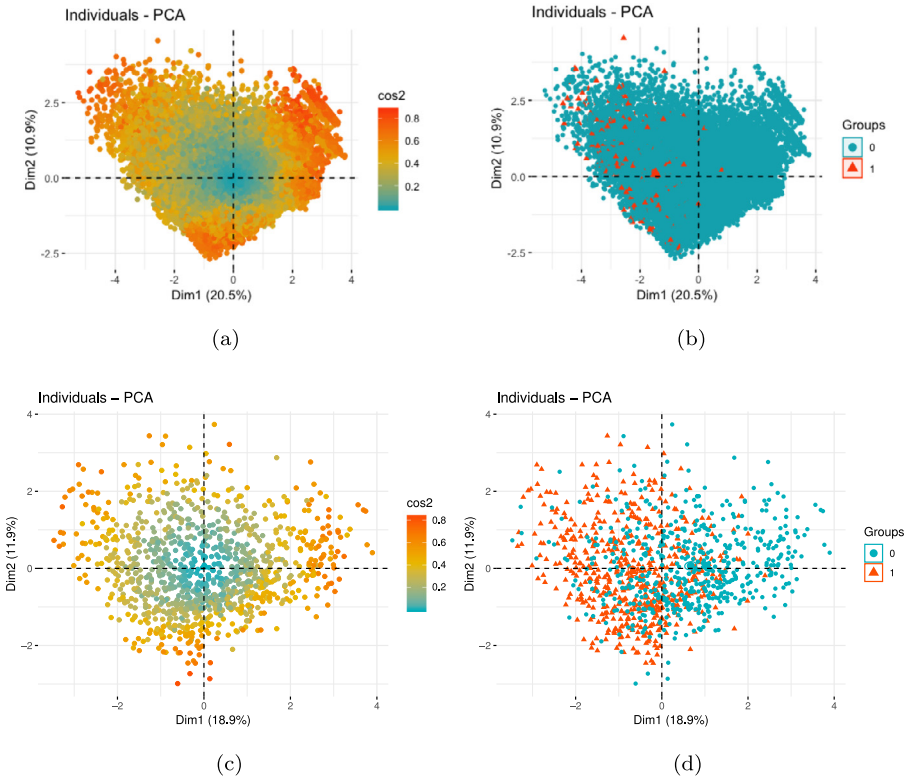


Fig. 8. Subspace representation of the different patient records in reference to the first two principal components. The observations are colour coded based on (a) *cos2* measure indicating the importance of principal components on the observation; and (b) status of stroke. We also use balanced dataset and observe the sub-space representation in (c) and (d).

We also compute the subspace representation of the different features for the balanced EHR dataset. We show it in Fig. 8(c) and (d). We observe that the observations with *stroke* and *no stroke* are scattered throughout the two principal axes. The observations with similar labels are not clustered together. Therefore, the features of the EHR records are important for efficient stroke prediction.

3.4. Discussion

This section discussed how principal component analysis can assist in a clear understanding of the original feature space of patient records. We showed that the first two principal components can cumulatively capture only 31.4% of the total variance in the input feature space. More so, first eight components can explain only about 88% of the total variance. Moreover, we studied the contribution of different patient attributes to the first two principal components. We could see that the two components do not represent the health records data perfectly. We also looked at the contribution of the first two principal components in the representation of individual health records. We could see that some health records can be represented by the two components, but some cannot. Thus, all principal components are needed to have a good representation of the variance in medical records. We cannot get a significant reduction of feature space for predictive modelling without a significant loss of variance in the data. Hence, we use all the principal components for predictive modelling of stroke occurrence.

In the next section, we compare the state of art machine learning classification techniques for predicting the occurrence of stroke in a patient's medical record. As discussed, we use ten patient attributes as input features to the models.

4. Stroke prediction

We provide a detailed analysis of various benchmarking algorithms in stroke prediction in this section. We benchmark three popular classification approaches — neural network (NN), decision tree (DT) and random forest (RF) for the purpose of stroke prediction from patient attributes. The decision tree model is one of the popular binary classification algorithm. This method involves building a tree-like decision process with several condition tests, and then applying the tree to the medical record dataset. Each node in this tree represents a test, and the branches correspond to the outcome of the test. The leaf nodes finally represent the class labels. The pruning ability of such algorithm makes it flexible and accurate, which is required in medical diagnosis. We also benchmark the dataset on random forest approach. The flexibility and ease of use of the random forest algorithm coupled with its consistency in producing good results, even with minimal tuning of the hyper-parameters makes this algorithm valuable in this application. The possibilities of over-fitting are limited by the number of trees existent in the forest. Moreover, random forest can also provide adequate indicators on the way it assigns significance to each of these input variables. We also benchmark the performance of a 2-layer shallow neural network. Artificial neural networks are quite popular these days, and they offer competitive results. We implement the feed-forward multi-layer perceptron model using the `nnet` R package.

4.1. Benchmarking using all features

Our dataset contains a total of 29072 medical records. Out of this, only 548 records belong to patients with stroke condition, and the remaining 28524 records have no stroke condition. This is a highly unbalanced dataset. This creates problem in using this data directly for training any machine-learning models. Therefore, we use random downsampling technique to reduce the adverse impact of the unbalanced nature of the dataset. We refer the 548 records as the minority class, and the remaining 28524 records with no stroke condition as the majority class. Subsequently, we create a dataset of 1096 observations, that consists of 548 minority samples and 548 majority samples. This balanced dataset is created by considering all the 548 minority samples,

Table 2
Performance evaluation of neural network, decision tree and random forest on our dataset of electronic medical records. We compare their performance for three different cases – (a) using all the original features, (b) using the PCA-transformed data of the first two principal components, and (c) using the PCA-transformed data of the first eight components. We choose 8 components, as 8 components are necessary to cumulatively contain more than 80% of the explained variance. We report the average value of precision, recall, F-score, accuracy, miss rate and fall-out rate, based on 100 experiments.

Features	Way	Precision	Recall	F-score	Accuracy	Miss rate	Fall-out rate
Original features (All)	DT	0.75	0.74	0.74	0.74	0.17	0.24
	RF	0.74	0.73	0.73	0.74	0.18	0.25
	NN	0.80	0.74	0.77	0.77	0.16	0.18
	CNN	0.74	0.72	0.73	0.74	0.17	0.24
	SVM	0.67	0.68	0.68	0.68	0.23	0.32
	LASSO	0.78	0.72	0.75	0.76	0.19	0.20
	ElasticNet	0.79	0.71	0.75	0.76	0.19	0.19
Original features (A+HD+AG+HT)	DT	0.78	0.71	0.74	0.75	0.20	0.21
	RF	0.76	0.74	0.75	0.75	0.18	0.24
	NN	0.78	0.71	0.74	0.75	0.19	0.20
PCA features (PC1 and PC2)	DT	0.78	0.65	0.71	0.73	0.24	0.19
	RF	0.71	0.68	0.69	0.69	0.23	0.28
	NN	0.77	0.67	0.72	0.74	0.22	0.20
PCA features (PC1 till PC8)	DT	0.75	0.68	0.72	0.73	0.21	0.23
	RF	0.73	0.69	0.71	0.72	0.21	0.25
	NN	0.80	0.68	0.73	0.75	0.21	0.17

Table 3
Accuracy variance for NN, SVM, LASSO and ElasticNet.

Features	Way	Precision	Accuracy variance
Original features (All)	NN	0.80	0.000377
	SVM	0.67	0.000470
	LASSO	0.78	0.000380
	ElasticNet	0.79	0.000467

Table 4
Hyperparameters of the CNN model.

Layers	In channels/Out channels	Kernel, Stride, Padding	Activation functions
Conv1	1/16	3, 1, 1	ReLu
Conv2	16/8	2, 1, 0	ReLu
Layers	In features/Out features	Kernel, Stride, Padding	Activation functions
Linear1	32/16	–	ReLu
Linear2	16/1	–	Sigmoid

and the remaining 548 majority samples are selected randomly from the 28524 patient records. All the three machine learning models are trained on this balanced dataset of 1096 observations.

In our experiment, another deep learning approach, the convolutional neural network (CNN) is implemented for the prediction of stroke. In our configuration, the number of hidden layers is four while the first two layers are convolutional layers and the last two layers are linear layers, the hyperparameters of the CNN model is given in Table 4. We use the same train and test split for CNN training and testing procedure, the ten inputs features are reshaped into 1 * 2 * 5 for inputs.

We also calculate accuracy variance for the benchmarking methods. Table 3 shows the experiments of accuracy variance for NN, SVM, LASSO and ElasticNet.

4.2. Benchmarking using top four features

Here we present results for stroke prediction when all the features are used and when only 4 features (A, HD, AG and HT) are used. These features are selected based on our earlier discussions. In addition to the features, we also show results for stroke prediction when principal components are used as the input. Since we observed that almost 8 principal components are needed to explain a variance of greater than 80%, we present results for both the cases when only first 2 principal components are used and when all the components are used. Table 2 shows the evaluation metrics for all the different configurations. In order to remove sampling bias, we perform 100 random downsampling experiments. The ratio of the number of training observations and testing observations is 70: 30. Table 2 reports the average values for all the approaches.

We observe that, among different machine learning approaches,[5] neural network works with better accuracy for different feature combinations. When we compare the neural network results for cases when all features are used and when only 4 features are used, we do not observe a significant improvement of all features over 4 features. Therefore, we can get a good stroke prediction accuracy of up to 78% with a low miss rate of 19% by using only 4 features (A, HD, AG and HT). This result might not be sufficient to guide treatment and prevention measures on an individual level but it will assist in supporting allocation and resources on a population and/or cohort level.

Here, for the case when the principal components are used as inputs, we observe that there is only slight improvement in accuracy of neural network, when all components are used compared to the first two components only. This is inline with our earlier discussion, where we illustrated that the first principal component can be important from the stroke prediction point of view. The variables that contributed most to the first component have shown good stroke prediction possibility. Therefore, use of only first two components have similar results compared to the case when all components are used.

When we compare the principal component results to the case when actual features are used, it can be observed that the accuracy of neural networks for both cases are comparable. However, if we look at the miss rate, the miss rates are higher for the case of principal components. The miss rate is an important evaluation metric as we would want to be able to detect all the strokes without a fail. We expect miss rate to be as low as possible and a slight degradation in the miss rate value is important for us for further consideration. Therefore, our analysis suggests that the best possible results for stroke prediction can be achieved by using neural network with 4 important features (A, HD, AG and HT) as input.

Finally, we illustrate the distribution of the accuracy values, by using the top 4 features — age, heart disease, average glucose level, hypertension from the dataset. We perform the experiments 100 times to remove any sampling bias in the training and testing sets. Fig. 9

[5] We do not benchmark our approaches against the McKinsey Kaggle challenge winner, as the model code is not publicly available.

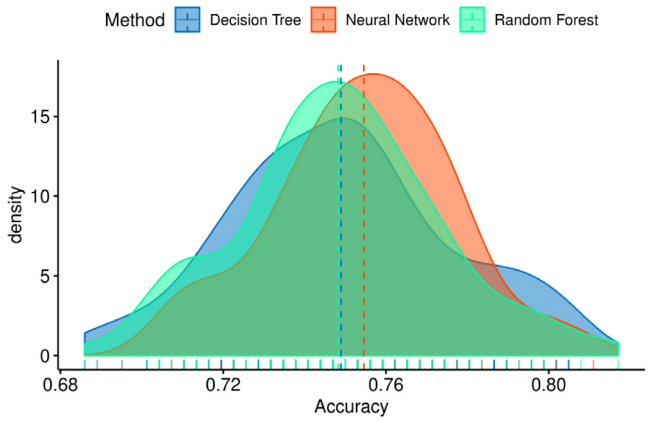


Fig. 9. Histogram distribution of classification accuracies obtained from the 100 experiments using top 4 features — age, heart disease, average glucose level, hypertension for the benchmarking algorithms.

illustrates this. We observe that most of them have similar performance, with their mean overlapping around similar values. We also compute the variance of the accuracies of the benchmarking methods for the 100 random sub-sampling observations. The variance of decision tree, neural network and random forest are 0.00073, 0.00049, and 0.00061 respectively. This indicates that there is no sampling bias involved in the benchmarking results.

5. Conclusion and future work

In this paper, we presented a detailed analysis of patients' attributes in electronic health record for stroke prediction. We systematically analysed different features. We performed feature correlation analysis and a step wise analysis for choosing an optimum set of features. We found that the different features are not well-correlated and a combination of only 4 features (A, HD, HT and AG) might have good contribution towards stroke prediction. Additionally, we performed principal component analysis. The analysis showed that almost all principal components are needed to explain a higher variance. The variable loadings however showed that the first principal component which has the highest variance might explain the underlying phenomenon of stroke prediction. Finally, three machine learning algorithms were implemented on a set of different features and principal components configurations. We found that neural network works the best with a feature combination of A, HD, HT and AG. The accuracy and miss rate for this combination are 78% and 19% respectively.

We have seen promising results from using just 4 features. The accuracy of the perceptron model cannot be improved further for primarily two reasons: lack of additional discriminatory feature set; and lack of additional dataset. We observed that most of the existing features in the EHR dataset are highly correlated to each other, and therefore do not add any additional information to the original feature space. Furthermore, a larger dataset will enable us to train our deep neural networks more efficiently. We plan to collect institutional data in our planned future work. The systematic analysis of the different features in the electronic health records will assist the clinicians in effective archival of the records. Instead of recording and storing all the features, the data management team can archive *only* those features that are essential for stroke prediction. Thus, in future, we plan to integrate the electronic records dataset with background knowledge on different diseases and drugs using Semantic Web technologies [22,23]. Knowledge graph technologies [23,24] can be used in order to publish the electronic health records in an interoperable manner to the research community. The added background knowledge from other datasets can also possibly improve the accuracy of stroke prediction models as well. We intend to collect our institutional dataset for further benchmarking of these machine learning methods for stroke prediction. We also plan to perform external validation of our proposed method, as a part of our upcoming planned work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

[1] G. Sivapalan, K. Nundy, S. Dev, B. Cardiff, J. Deepu, ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors, IEEE Transactions on Biomedical Circuits and Systems (2) (2022).
[2] H.C. Koh, G. Tan, et al., Data mining applications in healthcare, J. Healthc. Inf. Manage. 19 (2) (2011) 65.
[3] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, L. Hua, Data mining in healthcare and biomedicine: a survey of the literature, J. Med. Syst. 36 (4) (2012) 2431–2448.
[4] J.F. Meschia, C. Bushnell, B. Boden-Albala, L.T. Braun, D.M. Bravata, S. Chaturvedi, M.A Creager, R.H. Eckel, M.S. Elkind, M. Fornage, et al., Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American heart association/American stroke association, Stroke 45 (12) (2014) 3754–3832.
[5] P. Harmsen, G. Lappas, A. Rosengren, L. Wilhelmsen, Long-term risk factors for stroke: twenty-eight years of follow-up of 7457 middle-aged men in goteborg, sweden, Stroke 37 (7) (2006) 1663–1667.
[6] C.S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, D. John, Predicting stroke from electronic health records, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 5704–5707.
[7] M.S. Pathan, Z. Jianbiao, D. John, A. Nag, S. Dev, Identifying stroke indicators using rough sets, IEEE Access 8 (2020) 210318–210327.
[8] R.S. Jeena, S. Kumar, Stroke prediction using SVM, in: Proc. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 600–602, http://dx.doi.org/10.1109/ICCICCT.2016.7988020.
[9] S.-M. Hanifa, K. Raja-S, Stroke risk prediction through non-linear support vector classification models, Int. J. Adv. Res. Comput. Sci. 1 (3) (2010).
[10] J.K. Luk, R.T. Cheung, S. Ho, L. Li, Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects, Cerebrovasc. Dis. 21 (4) (2006) 229–234.
[11] S.N. Min, S.J. Park, D.J. Kim, M. Subramaniyam, K.-S. Lee, Development of an algorithm for stroke prediction: a national health insurance database study in Korea, Eur. Neurol. 79 (3–4) (2018) 214–220.
[12] M.S. Singh, P. Choudhary, Stroke prediction using artificial intelligence, in: 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), IEEE, 2017, pp. 158–161.
[13] P. Chantamit-o, Prediction of stroke disease using deep learning model.
[14] A. Khosla, Y. Cao, C.C.-Y. Lin, H.-K. Chiu, J. Hu, H. Lee, An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 183–192.
[15] C.-Y. Hung, C.-H. Lin, T.-H. Lan, G.-S. Peng, C.-C. Lee, Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database, PLoS One 14 (3) (2019) e0213007.
[16] D. Teoh, Towards stroke prediction using electronic health records, BMC Med. Inform. Decis. Mak. 18 (1) (2018) 1–11.
[17] C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, C.-C. Lee, Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 3110–3113.

[18] X. Li, H. Liu, X. Du, P. Zhang, G. Hu, G. Xie, S. Guo, M. Xu, X. Xie, Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation, in: AMIA Annual Symposium Proceedings, 2016, American Medical Informatics Association, 2016, p. 799.

[19] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowl.-Based Syst. 98 (2016) 1–29.

[20] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (1) (2017) 198–208.

[21] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[22] B. Tilahun, T. Kauppinen, C. Keßler, F. Fritz, Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation, JMIR Med. Inform. 2 (2) (2014).

[23] F. Orlandi, A. Meehan, M. Hossari, S. Dev, D. O'Sullivan, T. AlSkaif, Interlinking heterogeneous data for smart energy systems, in: 2019 International Conference on Smart Energy Systems and Technologies (SEST), IEEE, 2019, pp. 1–6.

[24] J. Wu, F. Orlandi, I. Gollini, E. Pisoni, S. Dev, Uplifting air quality data using knowledge graph, in: 2021 Photonics & Electromagnetics Research Symposium (PIERS), IEEE, 2021, pp. 2347–2350.