Fırat Tıp Dergisi/Firat Med J 2024; 29(4): 191-195

Clinical Research

Evaluation of Radial Basis Function Network and Supervised Machine Learning Methods on Brain Stroke Prediction Datasets

Kübra Elif AKBAŞ[1,a], Betül DAĞOĞLU HARK[1]

[1]*Fırat University Faculty of Medicine, Department of Biostatistics, Elazığ, Türkiye*

ABSTRACT

Objective: Supervised machine learning algorithms and neural networks are widely used classification methods in data mining. In this study, RBFN, one of the widely used supervised machine learning (SML) algorithms and neural network methods, was used according to the factors affecting the diagnosis of cerebral palsy, and it was aimed to evaluate their classification performance.

Material and Method: The dataset is an open source dataset, and there are a total of 4981 people with and without stroke. This dataset is modeled with RBFN from neural networks with four algorithms commonly used in supervised machine learning decision tree (DT), random forest (RF), and K-nearest neighbor (K-NN) and support machine vector (SVM). Their performance was evaluated according to performance criteria.

Results: The algorithms with the highest performance according to the accuracy criteria are DT (0.954), SVM (0.954), RBFN (0.954) and RF (0.953), respectively. The K-NN algorithm was found to be higher than other methods in terms of precision (0.061) and sensitivity (0.080).

Conclusion: The performances of DT, RF, SVM and RBFN methods were found to be close to each other in terms of accuracy criteria. In the decision-making process, the correct classification performance of these four methods is higher than K-NN.

Keywords: *Data Mining, Supervised Machine Learning, Neural Network, Performance Measures, Brain Stroke.*

ÖZ

Beyin Felci Verisetinin Tahmininde Radyal Temelli Fonksiyon Ağı ve Denetimli Makine Öğrenimi Yöntemlerinin Karşılaştırılması

Amaç: Denetimli makine öğrenmesi algoritmaları ve sinir ağları, veri madenciliğinde yaygın olarak kullanılan sınıflama yöntemlerindendir. Bu çalışmada, beyin felci hastalığının tanısının konulmasını etkileyen faktörlere göre yaygın olarak kullanılan denetimli makine öğrenmesi algoritmaları ve sinir ağları yöntemlerinden RBFN kullanılmış ve sınıflandırma performanslarının değerlendirilmesi amaçlanmıştır.

Gereç ve Yöntem: Veri seti açık kaynaklı bir veri seti olmak üzere inme ve inme olmayan toplam 4981 kişi yer almaktadır. Bu veri seti denetimli makine öğrenmesinde sıklıkla kullanılan dört algoritma (karar ağacı (DT), rastgele orman (RF), ve K-en yakın komşuluk (K-NN) ve destek makine vektörü (SVM) ile sinir ağlarından RBFN ile modellenmiştir. Ayrıca algoritma performansları performans kriterlerine göre değerlendirilmiştir.

Bulgular: Doğruluk kriterine göre performansı en yüksek olan algoritmalar sırasıyla DT (0,954), SVM (0,954), RBFN (0,954) ve RF(0,953) 'dür. K-NN algoritması ise kesinlik (0,061) ve duyarlılık (0,080) bakımından diğer yöntemlere göre daha yüksek bulunmuştur.

Sonuç: DT, RF, SVM ve RBFN metodları performansları doğruluk kriteri bakımından birbirlerine yakın bulunmuştur. Karar verme sürecinde bu dört yöntemin K-NN e göre doğru sınıflama performansı daha yüksektir.

Anahtar Sözcükler: *Veri Madenciliği, Denetimli Makine Öğrenmesi, Sinir Ağları, Performans Ölçütleri, Beyin Felci.*

Bu makale atıfta nasıl kullanılır: Akbaş KE, Dağoğlu Hark B. Beyin Felci Verisetinin Tahmininde Radyal Temelli Fonksiyon Ağı ve Denetimli Makine Öğrenimi Yöntemlerinin Karşılaştırılması. Fırat Tıp Dergisi 2024; 29(4): 191-195.

How to cite this article: Akbas KE, Dagoglu Hark B. Evaluation of Radial Basis Function Network and Supervised Machine Learning Methods on Brain Stroke Prediction Datasets. Firat Med J 2024; 29(4): 191-195.

ORCID IDs: K.E.A. 0000-0002-2804-000X, B.D.H. 0000-0002-5189-1929.

Stroke is one of the important health problems encountered today. Stroke, also known as a cerebrovascular accident, is a neurological disease caused by ischemia or bleeding of the cerebral arteries. This disease causes cognitive disorders. The two leading causes of death and disability worldwide are ischemic heart disease and stroke (1). In addition, the cost of hospitalization increases due to stroke. For these reasons, diagnosis, treatment, estimation of the clinical process, recommendation of therapeutic interventions and rehabilitation programs for stroke have become an important need. Machine learning (ML) is an important decision-making process in this process (2).

Advances in computer and information technologies have made the emergence of big data sets an inevitable situation. As the amount of data increases, uncovering patterns and trends and understanding data (i.e. learning from data) has become an important issue. An important area known as "machine learning" has developed to process big data statistically. ML is mainly based on supervised and unsupervised learning. Supervised machine learning (SML) assumes that the machine will learn from the data when a target variable is specified. Unsupervised learning is the opposite of supervised learning. Here, no target value or label is specified for the data. In addition, the data is visualized in two or three dimensions (3).

Radial basic function neural networks (RBFN) are one

[a]Yazışma Adresi: Kübra Elif AKBAŞ, Fırat University Faculty of Medicine, Department of Biostatistics, Elazığ, Türkiye
Tel: 0424 233 3555 e-mail: keet@firat.edu.tr
Geliş Tarihi/Received: 27.01.2023 Kabul Tarihi/Accepted: 29.05.2023

of the basic categories of neural networks (NN). The main architectures, learning techniques and applications of RBFN have been demonstrated in many studies (4-6). The learning and generalization abilities of these networks are very good. In particular, the learning rates are significantly faster compared to other multilayer neural networks (7).

The aim of this article is to compare the performances of ML and NN approaches, which are predictive methods for the diagnosis of stroke, which causes death and disability. Decision tree (DT), random forest (RF), k-nearest neighbors (K-NN) and support vector machines (SVM) are used as supervised machine learning (SML) methods, while RBFN was used as NN. First of all, SML and RBFN methods were compared and then these methods were evaluated within themselves.

MATERIAL AND METHOD

Dataset

The dataset includes a total of 4981 people with and without stroke, and the open access dataset was obtained from the relevant source https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset (8). In this data set, there are 10 explanatory variables and 1 response variable indicating the presence of stroke. While the total number of patients with stroke was 248 (5.0%), the number of patients without stroke was 4733 (95.0%). The explanation of the variables of the data set is given in table 1.

Table 1. *The detailed explanation of the variables in the dataset.*

Abbreviation	Explanation	Type	Role
Gender	(1 – Male, 2 – Female)	Categorical	Input
Age	Age (year)	Continuous	Input
Hypertension	(0 – No, 1 – Yes)	Categorical	Input
Ever Married	(0 – No, 1 – Yes)	Categorical	Input
Work Type	(1 – Private, 2 – Self-employed, 3-- Government Job, 4-- Children)	Categorical	Input
Residence type	(1 – Urban, 2 – Rural)	Categorical	Input
Avg glucose level	Average glucose level in blood	Continuous	Input
Smoking status	(1 – Smokes, 2 – formerly smoked, 3—never smoked)	Categorical	Input
BMI	Body mass index	Continuous	Input
Stroke	(0 – No, 1 – Yes)	Categorical	Output

Data Preprocessing

The data set was first evaluated in terms of extreme values for all variables. Then, the presence of missing data was checked and it was determined that 1500 (30.1%) data were missing for the smoking variable. Due to the high amount of loss in this variable, this variable was assigned with the multiple imputation method.

Supervised Machine Learning Algorithms

Decision Tree (DT)

Decision trees are one of the powerful ML methods widely used in various fields such as image processing and pattern identification (9). The algorithm, as the name suggests, consists of a tree structure with root node, branches and leaf nodes displaying attributes, conditions and results respectively (10). Each node represents a property in a classification category, and each subset specifies a value that the node can access (9). The DT is a sequential model that efficiently and harmoniously combines a set of core tests in which a numerical feature is compared with a threshold value in each test. Conceptual rules are much easier to construct than numerical weights in a NN between nodes. DT, which is mainly used for grouping in data mining, is a model that is also used for classification purposes (9).

Random Forest (RF)

The random forest (RF) method was first proposed by Breiman. This algorithm includes many decision trees. This method can be used in both regression and classification problems. Also, RF is one of the best ML algorithms that can be applied to many different fields. This algorithm consists of two parts. The first of these is training data. The second is validation data. From these two data sets, many random decision trees are created with bootstrap samples. The branching of each tree is determined by randomly selected estimators at the nodes. The final estimate of RF is the mean of all results from each tree. Each tree weights are taken into account while estimating the RF in the algorithm, and therefore each tree is not examined individually (11).

K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) are one of the most common classification techniques used in machine learning. K-NN is considered a non-parametric method when data distribution assumptions are not met. K-NN takes into account the equivalence of the new data with the existing data and places the new data in the nearest existing class. K-NN is used in recognition problems as well as regression problems (12).

Support Vector Machines (SVM)

Support vector machines (SVM) is a data-driven machine learning approach that deals with assigning class labels to unlabeled data. It is a predictive binary classification procedure. SVM is based on the maximum margin function that divides the observations into two classes. This function divides the data into two classes using the set of observations with known labels. The new unlabeled data is then assigned a class based on the classifier function and their geometric position (13). Some datasets are not linearly separable and any dividing line, no matter how narrow the margin, can cause misclassification (14). This problem can be solved by using the softer margin to estimate training samples with an acceptable misclassification (15).

The disadvantage of support vector machines is that the classification result is binary and the probability of class membership is not estimated.

Radial Basis Function Network (RBFN)

The radial basis function network (RBFN) is used for classification and is a feed forward NN structure with a single hidden layer. The term 'feed-forward' means that

neurons are organized in layers in a layered NN. This NN consists of three layers. These are the input layer, hidden layer and output layer. The input layer consists of input data. The hidden layer transforms the data from the input field to the hidden field using a non-linear function. The linear output layer gives the response of the network. The Euclidean distance between the input vector and the center of each hidden unit in an RBFN is determined by the argument of the activation function of that unit (16, 17).

Data Analysis

IBM SPSS Statistics Version 22.0 package program was used for statistical analysis of the data. Categorical input variables were summarized as frequency and percentage, and continuous input variables as mean and standard deviation. Categorical input variables and the presence of stroke, which is the output variable, were compared using the Chi-Square test statistic. The mean difference between the presence of stroke and continuous input variables was statistically tested using the independent samples t-test. Statistical significance level was taken as 0.05 in all tests.

R-studio software language was used for the application and comparison of machine learning and deep learning methods.

RESULTS

Descriptive statistics and p values of output variables explaining the presence of stroke are given in table 2.

Table 2. *Descriptive statistics of patients.*

Variables		Stroke		p value
		No (n =4733) mean±sd	**Yes (n =248) mean±sd**	
Gender*	**Male Female**	1966(41.5) 2767(58.5)	108(43.5) 140(56.5)	0.552
Age		42.141∓22.345	67.820∓12.671	**<0.001**
Hypertension*	No Yes	4320(91.3) 413(8.7)	182(73.4) 66(26.6)	**<0.001**
Ever Married*	No Yes	1672(35.3) 3061(64.7)	29(11.7) 219(88.3)	**<0.001**
Work Type*	Private Self-employed Gov-Job Children	2712(57.3) 739(15.6) 611(12.9) 671(14.2)	148(59.7) 65(26.2) 33(13.3) 2(0.8)	**<0.001**
Residence type*	Urban Rural	2397(50.6) 2336(49.4)	135(54.4) 113(45.6)	0.268
Avg glucose level		104.569∓43.602	132.176∓61.771	**<0.001**
Smoking status*	Smokes formerly smoked never smoked	1044(22.1) 1147(24.2) 2542(53.7)	61(24.6) 84(33.9) 103(41.5)	**<0.001**
BMI		28.410∓6.834	30.187∓5.658	**<0.001**

**n (%) is used as descriptive statistics.*

The difference in patients with stroke in Table 2 according to age, presence of hypertension, ever married, type of employment, mean glucose level, smoking and BMI variables was found to be statistically significant (p value <0.001 for each variable). The stroke outcome variable for gender and residence type did not differ significantly (p values 0.552 and 0.268, respectively).

According to the findings from table 3, the DT algorithm has the highest performance in terms of accuracy, F_1 score and specificity.

Table 3. *Results obtained according to classifier performance criteria of ML and RBFN algorithms.*

Algorithms	Classifier performance measures				
	Accuracy	**F_1 score**	**Precision**	**Sensitivity**	**Specificity**
DT	**0.954**	**0.977**	NaN	0.000	**1.000**
RF	0.953	0.976	0.000	0.000	0.999
K-NN	0.902	0.948	**0.061**	**0.080**	0.941
SVM	**0.954**	0.976	0.000	0.000	0.999
RBFN	**0.954**	0.976	0.000	0.000	0.999

**The algorithm with the highest value for accuracy, F1 score, precision, sensitivity and specificity is the best algorithm.*

The accuracy, F_1 score and specificity performance criteria for SVM, RF and RBFN algorithms were close to the DT algorithm. The K-NN algorithm is higher than other algorithms in terms of precision and sensitivity.

DT, SVM, RBFN and RF algorithms are very close to each other in terms of performance criteria. Accuracy; calculated based on true positive and true negative observations, F_1 score; calculated based on false positive and false negative observations, and specificity; calculated based on negative estimates within the true negative. Therefore, DT, SVM, RBFN, and RF algorithms performed similarly in terms of performance measures involving negative estimates. While precision is the rate of correct classification within a positive prediction, sensitivity is the rate of correct classification within a true positive. Accordingly, the K-NN algorithm is a more precise method for accurate positive classification.

DISCUSSION

Stroke disease occurs when there is a blood flow disorder or deficiency in various parts of the brain. This causes the cells in the damaged areas of the brain to not receive the nutrients and oxygen they need, and as a result, the cells die. Stroke is a medical emergency that requires immediate medical attention. It is important to identify early diagnosis and appropriate treatment management in order to prevent the harm and damage it will cause (18). Stroke is a serious and common disease. At the same time, stroke, many people today suffer from acute brain attacks, which are ischemic strokes caused by blood clots blocking the cerebral arteries (19). Therefore, stroke is a critical medical condition that must be treated before it worsens.

Brain stroke can be detected early. Early detection can reduce the effects of this stroke on the brain and other parts of the body. The aim of our work is to detect the presence of the disease early with machine learning algorithms and RBFN when dealing with a medical diagnostic field such as cerebral palsy. ML and RBFN applications have become very common especially in the medical field. It provides great convenience to physicians in making clinical decisions and predictions and in the decision-making process of the disease. In addition, ML and RBFN solutions are needed to cope with the limited number of doctors and the rapidly increasing challenges of large data sources.

In their study, Singh et al. (20) used five different ML techniques to predict stroke in the "Cardiovascular Health Study (CHS)" dataset. In their studies, C4.5 algorithm, DT, principal component analysis (PCA), artificial neural networks (ANN) and SVM methods were used. Kivrak et al. (21) predicted mortality using machine learning approaches such as RF, K-NN, extreme gradient boosting (XGBoost) and deep learning on an open-source COVID-19 dataset. Sailasya and Kumari (22) took various physiological factors and used machine learning algorithms such as logistic regression (LR), RF, DT, K-NN, SVM, and naive bayes (NB) to train five different models to accurately predict the probability of stroke in the brain.

In this article, DT, RF, K-NN, SVM and RBFN are the algorithms used in the decision making process. DT, RF, SVM and RBFN methods were determined as the methods with the best performance.

Cicek and Kücükakcalı (5) compared the multilayer perceptron neural network (MLPNN) and RBFN methods using the Prostate Cancer Data Set in their study and found that the MLPNN method was more successful in classification. Tan et al. (6) compared RBFN and SVM methods in their study. They did not find a significant difference between the method results in the study.

The article has limitations. One of them is that the data set has an unbalanced structure in terms of output variable. So the positive observation rate is quite low. Therefore, the outputs of the algorithms for positive classification are low. In other studies, it is recommended to compare ML and RBFN algorithms considering this unbalanced structure.

Conclusion

Stroke is a critical medical condition that needs to be treated before it gets worse. Thanks to ML and RBFN, stroke can be predicted early and its serious side effects can be reduced. By using these systems, medical doctors can diagnose cerebral palsy earlier and thus take the necessary measures to reduce the effect of stroke. To facilitate this decision-making process, the use of DT, RF, SVM and RBFN algorithms is suggested in the article.

Conflicts of Interest

The authors declare that there is no conflict of interest.

REFERENCES

1. Johnson W, Onuma O, Owolabi M, Sachdev S. Stroke: a global response is needed. Bulletin of the World Health Organization. 2016; 94: 634.
2. Sirsat MS, Fermé E, Câmara J. Machine learning for brain stroke: a review. J. Stroke Cerebrovasc. Dis 2020; 29: 105162.
3. Muthukrishnan R, Rohini R, editors. LASSO: A feature selection technique in predictive modeling for machine learning. 2016 IEEE international conference on advances in computer applications (ICACA); 2016: IEEE.
4. Kaya MO. Computer-aided model for the classification of acute inflammations via radial-based function artificial neural network. The J Cogn Syst 2021; 6: 1-4.
5. Cicek İB, Kücükakcalı Z. Classification of prostate cancer and determination of related factors with different artificial neural network. Middle Black Sea J Health Sci 2020; 6: 325-32.
6. Tan X-h, Bi W-h, Hou X-l, Wang W. Reliability analysis using radial basis function networks and support vector machines. Computers and Geotechnics 2011; 38: 178-86.
7. Pedrycz W. Conditional fuzzy clustering in the design of radial basis function neural networks. IEEE transactions on neural networks. 1998; 9: 601-12.
8. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.
9. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. JASTT 2021; 2: 20-8.
10. Afrin S, Shamrat FJM, Nibir TI et al. Supervised machine learning based liver disease prediction approach with LASSO feature selection. BEEI 2021; 10: 3369-76
11. Yeşilkanat CM. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals 2020; 140: 110210.
12. Ghosh P, Azam S, Jonkman M et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access 2021; 9: 19304-26.
13. Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. Eur J Oper Res 2018; 265: 993-1004.
14. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nat Methods 2018; 15: 5.
15. Maglogiannis IG. Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies: Ios Press; 2007.
16. Qiu-Hao H, Yun-Long C. Assessment of karst rocky desertification using the radial basis function network model and GIS technique: a case study of Guizhou Province, China. Environ Geol 2006; 49: 1173-9.
17. Shah MH, Dang X. Classification of spectrally efficient constant envelope modulations based on radial basis function network and deep learning. IEEE Communications Letters 2019; 23: 1529-33.
18. Tazin T, Alam MN, Dola NN, Bari MS, Bourouis S, Monirujjaman Khan M. Stroke Disease Detection and Prediction Using Robust Learning Approaches. J Healthc Eng 2021; 2021: :7633381.
19. Alexopoulos E, Dounias G, Vemmos K. Medical diagnosis of stroke using inductive machine learning. Machine Learning and Applications. Machine Learn Med Applicat 1999: 20-3.
20. Singh MS, Choudhary P, Thongam K, editors. A comparative analysis for various stroke prediction techniques. International Conference on Computer Vision and Image Processing; 2019: Springer.
21. Kivrak M, Guldogan E, Colak C. Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. Comput Meth Prog Bıo 2021; 201: 105951.
22. Sailasya G, Kumari GLA. Analyzing the performance of stroke prediction using ML classification algorithms. Int J Adv Comput Sci Appl 2021; 12: 539-45.