DOI 10.2478/ama-2019-0026

ATTRIBUTE SELECTION FOR STROKE PREDICTION

Małgorzata ZDRODOWSKA*

*Faculty of Mechanical Engineering, Department of Biocybernetics and Biomedical Engineering
Bialystok Technical University, ul. Wiejska 45C, 15-351 Bialystok, Poland

m.zdrodowska@pb.edu.pl

received 24 April 2019, revised 28 September 2019, accepted 30 September 2019

Abstract: Stroke is the third most common cause of death and the most common cause of long-term disability among adults around theworld. Therefore, stroke prediction and diagnosis is a very important issue. Data mining techniques come in handy to help determine the correlations between individual patient characterisation data, that is, extract from the medical information system the knowledge necessary to predict and treat various diseases. The study analysed the data of patients with stroke using eight known classification algorithms (J48 (C4.5), CART, PART, naive Bayes classifier, Random Forest, Supporting Vector Machine and neural networks Multilayer Perceptron), which allowed to build an exploration model given with an accuracy of over 88%. The potential features of patients, which may be factors that increase the risk of stroke, were also indicated.

Keywords: data mining, classifier, J48 (C4.5), CART, PART, naive Bayes classifier, Random Forest, Support Vector Machine, Multilayer Perceptron, haemorrhagic stroke, ischaemic stroke

1. INTRODUCTION

According to the American Heart Association and the American Stroke Association, stroke is a sudden, focal, vascular damage to the central nervous system (brain, retina or spinal cord), whose condition is to confirm the presence of a stroke in neuroimaging or persistence and focal symptoms over 24 hours, while excluding other causes of neurological disorders. Within the meaning of this definition, the stroke diagnosis also includes patients with infarction-related focuses revealed in imaging studies whose clinical symptoms have resolved in less than 24 hours (Sacco et al., 2013).

Stroke is a sudden life-threatening condition, which requires hospitalisation. Stroke is caused by a sudden disturbance of the blood supply to the brain. This is the case if a large artery that supplies blood to the brain or a small intracerebral artery will close, severely narrow or crack and will not supply the blood with oxygen and nutrients to a specific area of the brain. The consequence of closing or major narrowing of the arteriole is an ischaemic stroke, otherwise known as cerebral infarction. In contrast, a haemorrhagic stroke (commonly referred to as 'stroke') occurs when the artery breaks and the blood spills over a certain area of the brain. Ischaemic strokes account for 85% of all strokes, whereas haemorrhagic strokes for 15%. Thus, the essence of stroke is acute cerebral insufficiency of various aetiologies, which causes reduced cerebral perfusion in the course of ischaemia or haemorrhage. Even a small ischaemic focal point located, for example, in the inner capsule can manifest itself in complete paralysis of the mid-body and loss of sensation in it (Mazur, 2005). Clinical signs of stroke include paresis or hemiplegia, halitic sensory disorder, aphasia speech disorders – the inability to say words and understand simple commands, visual disturbances – one-eye distraction, visual field disorders, dizziness and headaches with centrifugal feeling, accompanied by nausea, vomiting, balance disorders or double vision. The most important risk factors for stroke are hypertension, heart disease, diabetes, dyslipidaemia and coagulation disorders (Jacobs and Sapers, 2011; Strepikowska and Buciński, 2009; Trochimczyk et al., 2017).

Stroke is a very important medical and social problem. It is the main cause of permanent disability and lack of independence in the group of adults. It is estimated that 15 million people fall for a stroke every year in the world, and about 5 million people die (Sacco et al., 2013). This disease is in third place among causes of death (after cardiovascular disease and cancer) (Mackay and Mensah, 2004). Long-term disability results in serious social and economic consequences of patients and their families. Economic aspects are related not only to the costs of hospital treatment but, above all, also to long-term care, long-term rehabilitation and long-term treatment of patients after stroke. Among patients who survived the stroke, as much as 60% remain less or less disabled. In this group, half of the patients are not dependent or require constant care (Mackay and Mensah, 2004). Since stroke is a serious threat to life, proper and fast help is very important. The earlier a patient with this disease goes to a hospital, the greater the chance of survival and the avoidance of severe disability (Strepikowska and Buciński, 2009; Trochimczyk et al., 2017).

Due to the development of data mining techniques, prediction and diagnosis of stroke are possible with increasing accuracy. Medical information systems have enormous data resources on patients, their health status and treatment processes. However, this information becomes meaningful only when their correlations with other information can be determined. This is what data mining algorithms deal with, which allow finding new, significant and correlated information in the data set. They are used especially to classify diseases, patients as well as for predictive pur-

poses (Maimon and Rokach, 2010; Yoo et al., 2012; Dardzińska, 2013; Alaiz-Moreton et al., 2018; Derlatka et al. 2019).

The aim of this study is to evaluate the effectiveness of selected data mining techniques for predicting the occurrence of stroke, as well as to identify those features of the patient that have the greatest impact on the occurrence of stroke. It will also be analysed how the reduction of attributes describing patients affects the accuracy of the classification.

2. METHODOLOGY

The study analysed the data of patients affected by stroke. The test sample included 215 patients diagnosed with haemorrhagic stroke and ischaemic stroke. The characteristics of the test sample are presented in Table 1.

Tab. 1. Characteristic of patients

	Amount	Average age (min–max)	Haemorrhagic stroke	Ischaemic stroke
Women	118	76, 47 (42–97)	15	103
Men	97	72, 15 (39–91)	16	81
All	**215**	**74, 53 (39–97)**	**31**	**184**

Each of the patients is characterised by 34 attributes (including age, sex, presence or absence of atrial fibrillation, blood pressure, level of consciousness, laboratory test results, types of drugs administered, etc). Part of the patients' data was incomplete.

Before data processing, the data were cleaned (incorrect values, incorrect entries, etc) and prepared for processing in the WEKA software. Columns and rows with large gaps were removed; data from non-significant assumptions were removed; gaps in numerical data were filled and data were discretised and normalised. In addition, the data were divided into two separate data sets: a set of training data (learning) to build the model (80%) and a set of test data for model evaluation (20%).

WEKA is software in the field of machine learning and knowledge acquisition created in the JAVA programming language environment, was used to analyse the data. The WEKA program, created at the University of Waikato in New Zealand, is a set of algorithms used to carry out data mining tasks, which allows, among others, for initial data processing, grouping, classification, regression, visualisation or the discovery of association rules (Witten et al., 2011).

In this work, classification was made for four models:

- Model 1: all available variables (34 attributes).
- Model 2: variables indicated in the literature as risk factors (11 attributes: hypertension, ischaemic heart disease, previous myocardial infarction, carotid artery stenosis, atrial fibrillation, diabetes, dyslipidaemia, smoking, alcohol, age and gender).
- Model 3: variables extracted using attributes selection – chi-square test (nine attributes: antiplatelet drugs, high-density lipoprotein (HDL), calcium (Ca)-blocker, low-density lipoprotein (LDL), diuretic, diabetes, ischaemic heart disease, total cholesterol and statin).
- Model 4: new variables – analysis of the principal component analysis (PCA) (five new attributes).

Due to the high accuracy, eight algorithms were used for the classification (Han and Kamber, 2006; Witten et al., 2011; Aggarwal, 2015; Frank et al., 2016, Chen et al., 2017; Kiranmai and Laxmi, 2018; Zdrodowska et al., 2018):

- J48 (C4.5) is an implementation of the C4.5 decision tree algorithm, which builds trees from a training set using entropy (information theory). It involves recursively visiting each decision-making node and selecting a possible division. To select the optimal division of the training set in the algorithm, the information gain is calculated. The C4.5 algorithm recursively visits each decision node, selecting a possible division, until further subdivisions are possible, uses trees that do not have to be binary, creates separate branches for each value of the qualitative attribute.
- CART is a very popular data classification method used to build decision trees. Its main features are high efficiency, the ability to build a tree both based on discrete and continuous data, creation of binary nodes (from each node leave at most two branches) and division of classes of solutions into superclasses (groups of classes). Like most algorithms, CART may interrupt its operation based on the interruption criterion. This criterion is determined based on the number of incorrect classifications and the number of tree leaves. It is this action that classifies the algorithm as a regression algorithm, as it evaluates and predicts the result besides the classification. There are several modifications of the CART algorithm; they differ mainly in the way of breaking the tree structure and assigning labels to the nodes.
- PART: the key of the PART algorithm is the construction of a partial decision tree, on the basis of which knowledge in the form of rules is discovered. The partial tree is an ordinary decision tree that undergoes construction and trimming operations until a stable subtree is found which cannot be simplified at a later stage. As soon as a partial tree is found, the rule is constructed and the tree discarded. This avoids rule generalisation and overdevelopment of the subtree, as happens when building rules with naive methods. Using the method of separation and winning in decision trees, the sensitivity and speed of the algorithm extracting the rule are increased. The algorithm does not require data optimisation.
- Naive Bayes Classifier is one of the machine learning methods used to solve sorting and classification problems. The task of the Bayes classifier is to assign a new case to one of the decision classes, while the set of decision classes must be finite and defined apriori. The naive Bayes classifier is a statistical classifier based on Bayes' theorem. In terms of efficiency, the naive Bayes classifier is comparable to classification algorithms by induction of decision trees and neural network classification methods. It is characterised by high accuracy and scalability even for very large data volumes. The naive Bayes classifier assumes that attribute values in classes are independent.
- Random Forest is a method of classification (and regression) involving the creation of multiple decision trees based on a random set of data. The idea of this algorithm is to build a team of experts from random decision trees, where, unlike classic decision trees, random trees are built on the principle that a subset of the analysed features in the node is randomly selected. In addition, individual trees from random forest trees are built according to the concept of bugging. Random forests are considered one of the best classification methods. Single

DOI 10.2478/ama-2019-0026

random forest classifiers are decision trees. The Random Forest algorithm is very well suited for trial testing, where the observation vector is a large dimension. Their additional advantage is the ability to use the learned random forest for other issues than just for classification. For example, based on trees from the forest, one can determine the ranking of variables, and thus determine which variables have better predictive properties.

- Support Vector Machine (SVM) offers very high accuracy compared with other classifiers, and its advantage is the use of non-linear data. The SVM can easily handle many continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes and generates the optimal hyperplane in a repetitive manner that is used to minimise the error. The basic idea of the SVM is to find the maximum boundary hyperplane (MMH) that best divides the data set into classes. SVM can be used for both classification and regression challenges. However, it is most often used in classification problems. The main purpose of the algorithm is to segregate a given data set in the best possible way. The distance between the nearest points is called the margin. The goal is to select a hyperplane that has the maximum possible margin between carrier vectors in a given data set.
- Multilayer perceptron (MLP) is a unidirectional multilayer neural networks; in other words, MLP networks are the most frequently described and most frequently used in practical applications of neural architecture. Their dissemination is related to the development of the algorithm of back error propagation, which enabled effective training of this type of network in a relatively simple manner. The multilayer neural network can approximate any complex and complex mapping. At the same time, the user does not have to know or assume any form of dependency in the sought after model and does not even need to ask himself whether any mathematical modelling of dependencies exists at all. This feature, combined with an independent method of learning a neural network, makes it an extremely useful and convenient tool for all kinds of applications related to forecasting, classification or automatic control.

The ACC (total accuracy) measure was used to assess the above classifiers. In order to test the accuracy of the constructed models, a matrix of errors (Kasperczuk et al., 2019) was used. ACC is the total efficiency of the classifier, which determines the probability of correct classification, that is, the ratio of correct classifications to all classifications. It is expressed by the equation (Bramer, 2016):

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

where (Bramer, 2016):

- TP (true positive) is the number of observations correctly classified to the positive class.
- TN (true negative) is the number of observations correctly classified to the negative class.
- FP (false positive) is the number of observations classified to a positive class when in fact they come from a negative class.
- FN (false negative) is the number of observations classified to the negative class when in fact they come from a positive class.

3. RESULTS AND DISCUSSION

The results of the classification of training and test sets for each of the proposed models are shown in Figs. 1–5.

For model 1 (Fig. 1), which use all attributes, in the case of a training set, the highest accuracy was achieved for the Random Forest classifier (99.42%). Equally, high accuracy (95.51%) was obtained for the MLP neural network algorithm. The SVM algorithm turned out to be the weakest; however, the accuracy is still quite high (over 86%).

For the test set, it is observable that all the algorithms work correctly, giving more than 80% accuracy. The highest accuracy was obtained using the Naive Bayes algorithm (88.38%), while the weakest results (however, still high – over 81%) were obtained using the J48 algorithm. It is easy to see in Fig. 1 that the differences in accuracy for the training and test set are small, which indicates a good model.

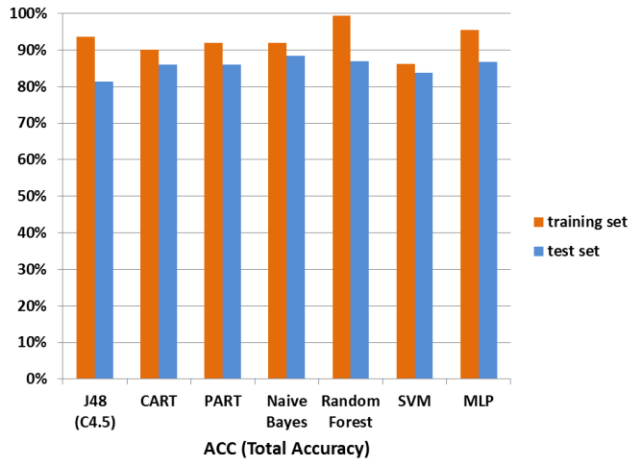

Fig. 1. Comparison of accuracy (ACC) of classifiers for model 1 (all variables) for training and test sets

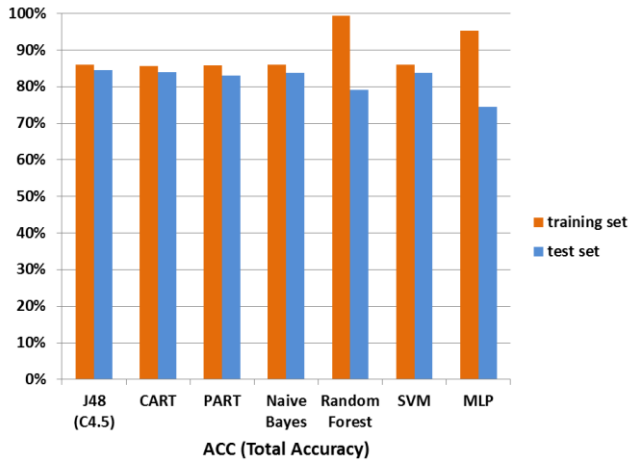

Fig. 2. Comparison of accuracy (ACC) of classifiers for model 2 (variables indicated in the literature as risk factors) for training and test sets

Let us look at model 2 (Fig. 2), which contains only variables indicated in the literature as risk factors (11 attributes: hypertension, ischaemic heart disease, past myocardial infarction, carotid stenosis, atrial fibrillation, diabetes, dyslipidaemia, smoking, alcohol, age and sex).

As we can see in Fig. 2, for model 2, the most accurate classifiers for the training set are Random Forest and MLP (99.42% and 95.35% accuracy, respectively). The total accuracy of the other classifiers is within 85–86%, which is quite a good result. For the test set, the J48 algorithm is the most accurate classifier (84.52%). Other classifiers also work correctly giving good matches. In principle, all algorithms (except Random Forest and MLP) give the test set a slightly lower accuracy index than in the case of the training set (about 2–3%), which indicates a very good model.

Let us move on to the model 3, which contains attributes extracted using feature selection – chi-square test (nine attributes: antiplatelet drugs, HDL, Ca-blocker, LDL, diuretic, diabetes, ischaemic heart disease, total cholesterol and statin). The classification results for this model are shown in Fig. 3.

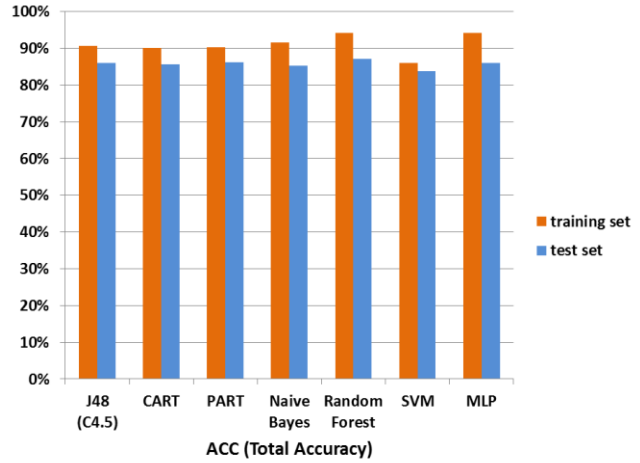

Fig. 3. Comparison of accuracy (ACC) of classifiers for model 3 (variables extracted after attributes selection) for training and test sets

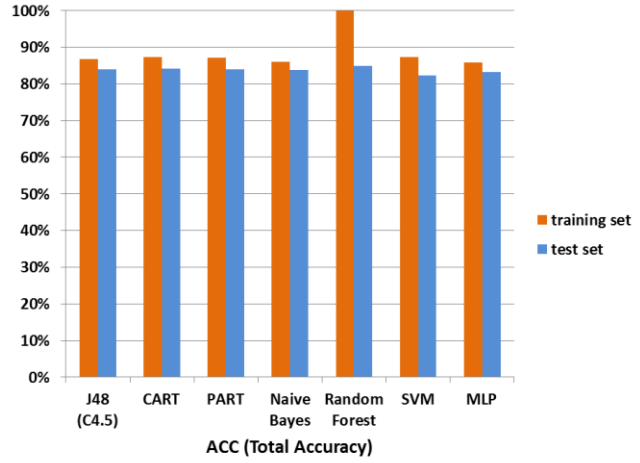

Fig. 4. Comparison of accuracy (ACC) of classifiers for model 4 (new variables – principal component analysis method) for training and test sets

In the case of model 3, the Random Forest and MLP algorithms proved to be the most accurate for the training set, which gave over 94% accuracy. Other algorithms also achieved accuracy above 90%. SVM turned out to be the weakest (86.05% accuracy). For the test set, the Random Forest algorithm was also the most accurate (87.05%), but for the other algorithms, high accuracy of over 83% was also achieved. It is worth adding here that in the case of the model with feature selection, the classification using the SVM achieved only a slightly lower accuracy for the test set than for the training set, which indicates a very good fit of the model (86.05% and 83.73%).

The last model is a model which contains five new variables obtained by PCA. The classification results for this model are shown in Fig. 5.

For model 4, the training set and the test set, the highest accuracy was obtained using the Random Forest algorithm. This accuracy is very high for both sets and reaches 84.85% for the test set, which shows us the benefits of PCA. The SVM algorithm proved to be the least accurate algorithm for model 4.

Fig. 4 shows that for all algorithms, the classification accuracy for the test set is slightly lower than for the training set, which indicates a good model.

Taking into consideration the comparison of the correctly classified objects in individual models (Fig. 5). The results obtained on the test set were compared because it shows the correct accuracy and usefulness of the model.

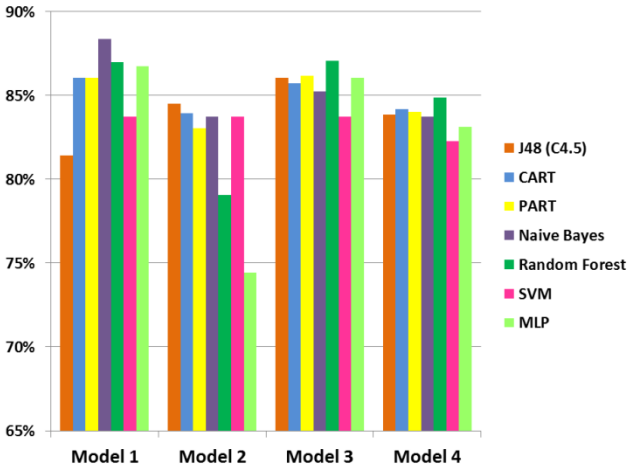

Fig. 5. Comparison of correctly classified objects in individual models (for test set)

The most accurate classification (88.38%) was obtained using the Naive Bayes algorithm for model 1, that is, working on all available attributes. A slightly lower accuracy (87.95%) was obtained for the Random Forest algorithm and model 3, which contains only attributes extracted using the feature selection – the chi-square.

The small difference between accuracy for the model containing all features and for models after feature selection. In several cases, classifiers for models after feature selection give better accuracy (or very similar) than models in which all attributes are included. It shows the usefulness of using attribute selection, which not only saves time but also gives better predictive results.

4. CONCLUSION

Knowledge discovery in databases is a dynamically developing field, whose rapid development is related to the growing number of databases and the size of information collected in them. Increasingly, data mining finds its application in medicine and medical information systems. The analysis of classification algo-

DOI 10.2478/ama-2019-0026

rithms carried out shows that they can be a very important tool supporting the prediction and diagnosis of various diseases. Detailed data analysis, their proper preparation and proper classification allow to achieve very accurate results. The study analyses the data of patients with stroke using eight popular classification algorithms, which allowed to build a mining model with an accuracy of over 87%.

By analysing the data of stroke patients, the difference between raw data and data after feature selection and analysis of basic components (PCA) was examined. The results obtained in this study confirmed the benefits of feature selection and analysis of basic components. The accuracy ratios of each of the algorithms used for the data after attribute reduction were similar to those with raw data.

Tests done with the help of WEKA software show that the attributes are most important in diagnosing a stroke. Important features were diabetes, ischaemic heart disease, dyslipidaemia, HDL, LDL and total cholesterol levels. It has also been shown that patients who have been given anticoagulants (antiplatelet drugs), blood lipid-lowering drugs (statins) and antihypertensive drugs (diuretic, Ca-blocker) have a higher risk of stroke. This may mean that patients who have been diagnosed with the above-mentioned diseases are more likely to have a stroke. This conclusion is confirmed in the literature, where hypertension, circulatory diseases and diabetes are mentioned as one of the main causes of stroke (Strepikowska and Buciński, 2009).

REFERENCES

1. **Aggarwal C.C.** (2015), *Data Classification Algorithms and Applications*, Chapman & Hall/CRC, New York.
2. **Alaiz-Moreton H., Fernández-Robles L., Alfonso-Cendón J., Castejón-Limas M., Sánchez-González L., Pérez H.** (2018),Data mining techniques for the estimation of variables in health-related noisy data, *Advances in intelligent systems and computing*, 649, 482–491.
3. **Bramer M.** (2016),*Principles of Data Mining*, Springer.
4. **Chen Y.C., Suzuki T., Suzuki M., Takao H., Murayama Y., Ohwada H.** (2017), Building a Classifier of Onset Stroke Prediction Using Random Tree Algorithm, International Journal of Machine Learning and Computing, 7(4), 61-66.
5. **Dardzińska A.** (2013), *Action Rules Mining*, Springer, Berlin.
6. **Derlatka M., Ihnatouski M., Jałbrzykowski M., Lashkovski V., Minarowski Ł.** (2019),Ensembling rules in automatic analysis of pressure on plantar surface in children with pes planovalgus,*Advances in Medical Sciences*, 64(1), 181-188.
7. **Frank E., Hall M.A., Witten I.A.** (2016), *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann.
8. **Han J., Kamber M.** (2006), *Data mining. Concepts and Techniques*, 2nd ed, Elsevier, San Francisco.
9. **Jacobs L.K., Sapers B.L.** (2011), Neurological Disease, In: Cohn S. (editor), *Perioperative Medicine*, Springer, London.
10. **Kasperczuk A., Daniluk J., Dardzińska A.** *(2019),* Smart Model to Distinguish Crohn's Disease from Ulcerative Colitis, *Applied Sciences*, 9(8), 1650.
11. **Kiranmai S.A., Laxmi J.A.** (2018), Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy, *Protection and Control of Modern Power Systems*, 3(29),https://doi.org/10.1186/s41601-018-0103-3.
12. **Mackay J., Mensah G.** (2004), *The Atlas of Heart Disease and Stroke: Global burden of stroke*, World Health Organization.
13. **Maimon O., Rokach L.** (ed). (2010), *Data mining and knowledge discovery handbook*, Springer.
14. **Mazur R., Świerkocka-Miastkowska M.** (2005), Stroke - first symptoms (in Polish), Choroby Serca i Naczyń, 2 (2), 84-87.
15. **Sacco R.L., Kasner S.E., Broderick J.P., Caplan L.R., Connors J.J., Culebras A., Elkind M.S., George M.G., Hamdan A.D., Higashida R.T., Hoh B.L., Janis L.S., Kase C.S., Kleindorfer D.O., Lee J.M., Moseley M.E., Peterson E.D., Turan T.N., Valderrama A.L., Vinters H.V.** (2013), An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association, Stroke, 44, 2064-2089.
16. **Strepikowska A., Buciński A.** (2009), Stroke – risk factors and prophylaxis (in Polish), *Farmakopea Polska*, 65(1), 46–50.
17. **Trochimczyk A.**, **Chorąży M., Snarska K.K.** (2017), An analysis of patient quality of life after ischemic stroke of the brain, *The journal of neurological and neurosurgical nursing*, 6(2), 44–54.
18. **Witten I.H., Frank E., Hall M.A.** (2011), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.
19. **Yoo I., Alafaireet P., Marinov M.** (2012), Data mining in healtcare and biomedicine, A survey of the literature, *Journal of the medical systems*, 35(4), 2431–2448.
20. **Zdrodowska M., Dardzińska M., Chorąży M., Kułakowska A.** (2018), Data Mining Techniques as a Tool in Neurological Disorders Diagnosis, *Acta Mechanica et Automatica*, 12(3), 217-220.

Acknowledgements: This work is supported by the Polish Ministry of Science and Higher Education of Poland under research project No. MB/WM/17/2018.