Iranian Journal of Operations Research
Vol. 11, No. 1, 2020, pp. 43-58
DOI: 10.29252/iors.11.1.43

Hybrid Method of Logistic Regression and Data Envelopment Analysis for Event Prediction: A Case Study (Stroke Disease)

Jafar Pourmahmoud[1,*], Maedeh Gholam Azad[2]

Predictive analytics is an area of statistics that deals with extracting information from data and using that to predict trends and behavioral patterns. Many mathematical models have been developed and used for prediction, and in some cases, they have been found to be very strong and reliable. This paper studies different mathematical and statistical approaches for events prediction. The main goal of this research is to design and construct a hybrid prediction method for events prediction, based on Logistic Regression (LR) method and Data Envelopment Analysis (DEA) technique. In this study, a novel hybrid algorithm was developed, and considering the kind of collected data, LR method was applied for input selection, and the capability of the additive (ADD) model of DEA was examined to predict the occurrence or non-occurrence of the events. To apply the proposed approach, the selected disease for the case study was a stroke. The results showed that any patient who was placed on the frontier has had a stroke by one or more risk factors. On the other hand, the observations that were not on the frontier had not suffered from a stroke. The overall accuracy of 88.5 percentages was obtained for the developed method.

***Keywords**: Data Envelopment Analysis, Logistic Regression, Additive Model, Risk Factor, Stroke Disease.*

The manuscript was received on 11/18/2019, revised on 01/04/2020, and accepted for publication on 03/21/2020.

1. Introduction

Any prediction method with relatively high accuracy can be a very useful tool in different industries. One of the most important fields that prediction can be used in, is the healthcare industry. Disease prediction has long been considered an important and challenging issue. Different approaches such as mathematical models, statistical methods, data mining, and multi-group classification methods were developed for classifying and analyzing data to be used in prediction. The mentioning approaches have been used for prediction in various fields including predicting and diagnosing the diseases in the healthcare industry, and some of these studies are presented in the next subsection.

[1] Department of Applied Mathematics, University of Azarbaijan Shahid madani, Tabriz, Iran,
Email: Pourmahmoud@azaruniv.ac.ir.
[*] Corresponding Author.
[2] Department of Applied Mathematics, University of Azarbaijan Shahid madani, Tabriz, Iran,
Email: m.gholamazad@azaruniv.ac.ir.

2. Literature Review

Dharani et al. [6], evaluated the performance of LR and SVR models to predict COVID-19 pandemic. Freeman et al. [8], presented the comparison of artificial neural networks (ANN) with LR in the prediction of in-hospital death after percutaneous transluminal coronary angioplasty. Fukunishi et al. [9], presented an Alzheimer-type of dementia prediction by sparse LR using claim data. Jee et al. [12], provided the stroke risk prediction model via COX graphs and logit function. Khanam and Y.Foo [15], compared two LR and SVM methods for diabetes prediction. Liew et al. [16], presented comparing ANN with LR in the prediction of gallbladder disease among obese patients. Lattanzi et al. [17], predicted the outcomes after stroke through LR analysis and discrimination and calibration tests. Mauthe et al. [18], studied predicting discharge destination of stroke patients using a mathematical model based on six items from the functional independence measure. Nguyen et al. [21], compared the prediction models for adverse outcomes in pediatric meningococcal disease using ANN and LR analyses. Nourijelyani et al. [22], presented the application of a mixed LR method in determination of effective factors related to visible goiter with health survey data. Nusinovici et al. [23], believed that LR was as good as machine learning for predicting major chronic diseases. Reed and Wu [28], used LR for risk factor modeling in stuttering research for stroke. Sergeev and Weckman [31], presented the prediction models using ANN and LR for examining cardiovascular disease treatment outcomes in patients with diabetes. Shukla et al. [32], applied LR method for predicting diabetes disease. Sposato et al. [33], examined the therapeutic decisions in atrial fibrillation for stroke prevention by LR. Changsheng Zhu et al. [34], investigated the improved LR method for diabetes prediction.

One of the most effective approaches used for prediction is DEA non-parametric approach, which was first proposed by Charnes et al. (CCR model) [4]. It allows multiple inputs and outputs to be used in a linear programming model for calculating the efficiency of decision-making units (DMUs). Subsequently, many models developed based on the CCR model, such as the BCC model (Banker et al.) [3], and ADD model (Carnes et al.) [5], etc.

Different studies are presented by various researchers in the field of prediction by DEA approach and were compared with data mining and statistical methods. For example, Alinezhad [1] presented a DEA model combined with Bootstrapping to assess the performance of one of the data mining Algorithms. He applied a two-step process for performance productivity analysis of insurance branches within a case study. Arasteh [2], presented a new method of combinatorial optimization considering stochastic values. Horváthová and Mokrišová [10], assessed the business financial health by formulating DEA model and verifying the estimation accuracy of this model in comparison with the logit model. Khalili Araghi et al. [13], evaluated the prediction power of DEA technique compared to logit and probit models in predicting corporate bankruptcy. Karamali et al. [14], examined the capability of ANN in a sensitivity analysis of the parameters of DEA model. Mendelova et al. [19], compared DEA and LR in corporate financial distress prediction. Mousavi et al. [20], provided a comparative analysis of two-stage distress prediction models. Premachandra et al. [24] applied DEA as a tool for bankruptcy assessment and compared the obtained results with LR method. Premachandra et al. [25], introduced the DEA as a tool for predicting corporate failure and success. Pourmahmoud and GholamAzad [26], presented a new model of DEA with binary data for prediction with the BIP-DEA model's name. Radovanovic et al. [29] proposed a two-phase approach with the combination of the DEA and machine learning for predicting the efficiency of NBA players. Silva E Souza [30], compared the prediction by LR with DEA prediction. Zhu et al. [35], presented combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies.

Based on the literature review, it was found that combined approaches of DEA and LR models are limited. If the parameters under evaluation for the occurrence or non-occurrence of an event are high, and no information of the most effective parameters is not available, a combination of LR and DEA approaches can be useful. For this purpose, in this study, we presented the hybrid algorithm for both approaches of LR and DEA. At the first stage, the effective parameters, and their relationships are selected by LR method using the proposed algorithm. At the second stage, the selected parameters are used in ADD model as inputs and the probabilities of the occurrence or non-occurrence of the events are examined. Finally, the classifying prediction table is offered. The proposed approach can be used in various fields, however, we used it for prediction in the field of healthcare for stroke. In this case study, 200 patients made up our study population. After the implementation of the proposed algorithm, the patients who have been placed on the frontier have had a stroke. Whereas, the patients who haven't been placed on the frontier have not had a stroke. To examine the accuracy of the algorithm, the classification table is presented. In this table, the probabilities of the correct classification rate and misclassification rate, and their errors are listed.

2.1. The Motivation of the Study and Existing Gaps

To have a deep understanding of the motivation of this study and to identify the existing gaps in the previous researches, we present a comparison between the current study and previous ones in Table 1.

Table 1. Comparison of the current study with previous researches

Researches	Research fields	Methods used
Previous researches [1, 10, 13, 19, 24, 25, and 30]	Business, financial health, bankruptcy, insurance branches, corporate failure, manufacturing.	BCC, ADD, SBM, and CCR models of DEA, and LR method.
Current study	Prediction of any events in various industries, health network, disease prediction, mortality, and survival rate prediction, etc.	Hybrid of the two ADD and LR models.
Comparison	Each of the methods used in previous research is unique to a specific industry. While the proposed approach in this study can be applied in an integrated manner in all fields.	

According to the literature review, using LR method in performance analysis has some drawbacks such as [19, 24, and 25]:

I. LR does not identify the individual inefficient units, and its parametric formulation property requires a pre-specified production function.
II. LR method needs a big volume of data to access the appropriate results in prediction.
III. LR method is not capable of ranking the criteria, and it is dependent on the number of observations for achieving reliable outcomes.

To overcome the above drawbacks, it is required that LR method to be combined with other nonparametric methods such as DEA approach, since, DEA approach is a valuable tool for performance measurement, and it does not need a large sample size of data for prediction [24, and 25].

All previous researches listed in the literature were used any DEA and LR approaches separately and they compared the obtained results together. While, in the current study, a new integrated framework will be proposed that combined ADD and LR models into an applicable algorithm. On

the other hand, these studies have usually been applied in bankruptcy. Whereas, not only can the proposed approach in this study be used in various industries but it can also be applied to the prediction of different diseases. For example, the case study of this research is predicting the occurrence of a stroke.
The framework of this study is as follows:
Section 2 presents the research methodology. Section 3 includes the proposed approach. Section 4 provides the case study and the results that are obtained. Finally, the conclusion is presented in section 5.

3. Methodology

The mathematical and statistical methods used in this study have been selected based on previous studies. Therefore, we have used LR method and ADD model of DEA approach with different assumptions and characteristics.

3.1. LR Method

When the subject study is a qualitative variable, LR method is usually used. In the regression discussion, the Y variable is considered a continuous variate. Sometimes, however, the Y variate is discontinuous, and in particular, has a dichotomous value: an outcome that either happens or does not. One example would be the survival of a premature infant related to birth weight, the presence or absence of a disease related to certain clinical or laboratory findings. Another example would be the reaction of an organism to medicine, whether a reaction is received or not. Therefore, Y can take on only one of two values, that is no (0) or yes (1). In such situations, LR method is recommended to be used. LR method maximizes the probability that an event may occur and uses the Chi-2 and Wald tests to examine the significance of the relationship. The Logit function is used as a link function in this method and its error follows a polynomial distribution. The general form of LR method is as follows [27]:

$$Logit(p) = Ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1,j} + ... + \beta_i x_{i,j}, \quad j = 1,...,N; \quad i = 1,...,m. \tag{1}$$

Where the $p = \Pr(y = 1)$.

$$p = \Pr(y = 1 | \vec{x}_i, \vec{\beta}) = \frac{e^{\beta_0 + \beta_1 x_{1,j} + ... + \beta_i x_{i,j}}}{1 + e^{\beta_0 + \beta_1 x_{1,j} + ... + \beta_i x_{i,j}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,j} + ... + \beta_i x_{i,j})}} \tag{2}$$

Where $x_{i,j}$ represents the predictor variables. $\beta_1, \beta_2, ...\beta_i$ Are estimating coefficients of the model for the independent variables and they are showing the effecting coefficient. The value of *p* shows the probability risk of happening or not of the events.

3.2. ADD Model of DEA

Mathematical models can be used throughout the prediction research process. Among different models based on mathematical methods, DEA approach was selected based on the type of events surveillance data. This model was selected according to the type of data and the fact that ADD model is translation-invariant in both inputs and outputs. ADD model was first introduced by

Charnes et al. (1985) [5]. Suppose that, there is a set of n units of DMUs, such that each $DMU_j\ (j=1,2,...,n)$ has m inputs and s outputs. The i^{th} input and r^{th} output of $DMU_j\ (j=1,2,...,n)$ are denoted by $x_{ij}\ (i=1,2,...,m)$ and $y_{rj}\ (r=1,2,...,s)$, respectively. ADD model evaluates the performance of a DMU_o, $o\in\{1,2,...,n\}$ (the DMU under evaluation) as follows:

$$\begin{aligned} \max\ \ & z=(\sum_{i=1}^{m}s_i^- + \sum_{r=1}^{s}s_r^+) \\ s.t.\ \ & \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^- = x_{io}, \quad i=1,...,m \\ & \sum_{j=1}^{n}\lambda_j y_{rj} - s_r^+ = y_{ro}, \quad r=1,...,s \\ & \lambda_j, s_i^-, s_r^+ \geq 0,\ j=1,...,n, i=1,...,m, r=1,...s. \end{aligned} \tag{3}$$

Where s_i^- and s_r^+ represent input and output slacks of the DMU_o under evaluation. The objective function of model (3) is the sum of slacks to be maximized. The DMU_o is efficient or lies on DEA frontier, if and only if $s_i^{*-}=s_r^{*+}=0$ at optimality. This indicates that the unit under evaluation is Pareto-efficient. In the context of the occurrence of an event, if the sum of the objective function is zero, the event occurs in the future with a high probability. Conversely, the event occurs in the future with low probability, if the sum of the objective function is greater than zero.

4. Proposed Approach

LR method is easy to implement and interpret. It also has good accuracy for many simple data sets and it performs well when the dataset is linearly separable. But, if the number of observations is less than the number of features, LR should not be used, otherwise, it may lead to overfitting. Moreover, it cannot use multiple inputs to produce multiple outputs. DEA approach does not have these weaknesses and it has overcome these challenges. Moreover, it is efficient in ranking. In this study to predict whether an event occurs or not, a two-stage process was used. As LR method is able to determine the effective risk factors, and ADD model can discriminate efficient and inefficient DMUs, in this study these two methods were combined.

Suppose that there are N samples or DMUs and each sample (DMU) includes the number of parameters, whose information is zero or one. The following characteristic function is used to determine the relevant information:

$$P_A(x_{ij})=\begin{cases}1, & x_{ij}\in A\\ 0, & O.W\end{cases}$$

Where "A" represents the set of parameters discussed and x_{ij} is an i[th] parameter related to the j[th] of the sample (the j[th] DMU), $i = 1,2,...,m$, $m = |A|$ and $j = 1,2,...,N$, N shows the number of samples or DMUs.

Consider the following algorithm.

- **The Hybrid Algorithm:**

Step 1: Import the set of all of the parameters (set of A).

Step 2: Specify the number of samples to be examined.

Step 3: Import the data of all samples.

Step 4: Run LR method.

Step 5: Check the *sig* of all parameters.

a) *if* $sig > 0.05 \Rightarrow$ The related parameter is eliminated.

b) *if* $sig \leq 0.05 \Rightarrow$ The related parameter is selected.

Step 6: Import inputs and their corresponding outputs based on the parameters that are selected from step 5 for all DMUs (samples).

Note: When running ADD model, the DMUs that have errors in the table of data registry must be removed.

Step 7: Run ADD model for the data table of step 6.

Step 8: Based on the results extracted from step 7 and using the values of the slacks, specify the DMUs that are placed on the frontier. If all of the slacks are zero, then DMU is on the frontier and an event has occurred. Otherwise, if at least one of the slacks is positive, DMU is not on the frontier, and the event has not occurred.

Step 9: Classify all of the DMUs into the following four groups: (1) The events ***(E)*** have had occurred and the DMUs are on the frontier ***(F)*** $(E \cap F)$, (2) The events ***(E)*** have had occurred and the DMUs aren't on the frontier ***(NF)*** $(E \cap NF)$, (3) The events ***(NE)*** haven't had occurred and the DMUs are on the frontier ***(F)*** $(NE \cap F)$, (4) The events ***(NE)*** haven't had occurred and the DMUs aren't on the frontier ***(NF)*** $(NE \cap NF)$.

Step 10: Compute the following probabilities for the above four groups:

(1): $P\left(\frac{(E \cap F)}{E}\right)$ = the number of events that have had occurred and are on the frontier, divided by the total number of events that occurred.

(2): $P\left(\frac{(E \cap NF)}{E}\right)$= the number of events that have had occurred and aren't on the frontier, divided by the total number of events that occurred.

(3): $P\left(\frac{(NE \cap F)}{NE}\right)$ = the number of events that have hadn't occurred and are on the frontier, divided by the total number of events that have hadn't occurred.

(4): $P\left(\frac{(NE \cap NF)}{NF}\right)$ = the number of events that have hadn't occurred and aren't on the frontier, divided by the total number of events that have hadn't occurred.

Interpret the results of the probabilities.

Step 11: Calculate the misclassification rate and the correct classification rate.

The percentage of the misclassification rate denoted by I_{MC} is as follows

$$I_{MC} = \left(\frac{(NE \cap F) + (E \cap NF)}{N}\right)$$

The percentage of the correct classification rate denoted by I_{CC} is as follows

$$I_{CC} = \left(\frac{(E \cap F) + (NE \cap NF)}{N}\right)$$

Where *N* is the number of samples.

Step 12: Stop.

The proposed approach can be used in various fields such as healthcare industry (i.e., occurrence or non-occurrence of the diseases), failure in industrials, bankruptcy in the firms, financial distress, and so on. The biggest advantage of this approach is that it can be applied with multiple inputs to produce multiple outputs.

5. Case Study: Stroke Disease

Today, healthcare organizations face increasing pressure to provide improved patient care. To do so, organizations are straightened to predictive analysis [8]. Selected as a case study stroke is the third leading cause of death and the serious cause of long-term disability in humans [12, 17]. Identifying the most effective risk factors for stroke is critical for health organizations. By correct recognition of these risk factors, it is possible to achieve effective prevention and treatment.

In this study, we aim to find out whether it is possible to predict stroke using the provided information by the patient or not. To achieve this goal, the study was performed on several patients. The case study population and the results of the implementation of the hybrid algorithm are analyzed as follows.

5.1. Study Population

Of 5411 patients with stroke, who have been referred to Imam Reza and Razi hospitals in Tabriz, East Azarbaijan Province, Iran, from April 2015 to April 2016, 200 patients were randomly selected and formed our study population [7].

5.2. Data Collection

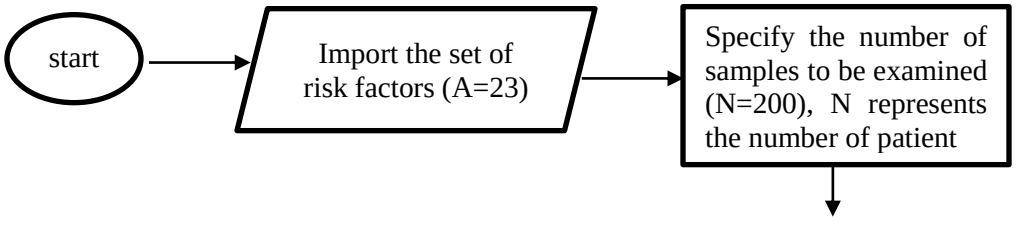
The gathered data from the patients are quantitative and qualitative. The dependent variable in this study is considered to be a stroke. After a literature review and consultation with neurology practitioners, 23 risk factors were selected as a potential reason or effecting reason on the stroke, shown in Table (2). These risk factors were also listed in clinical reports.

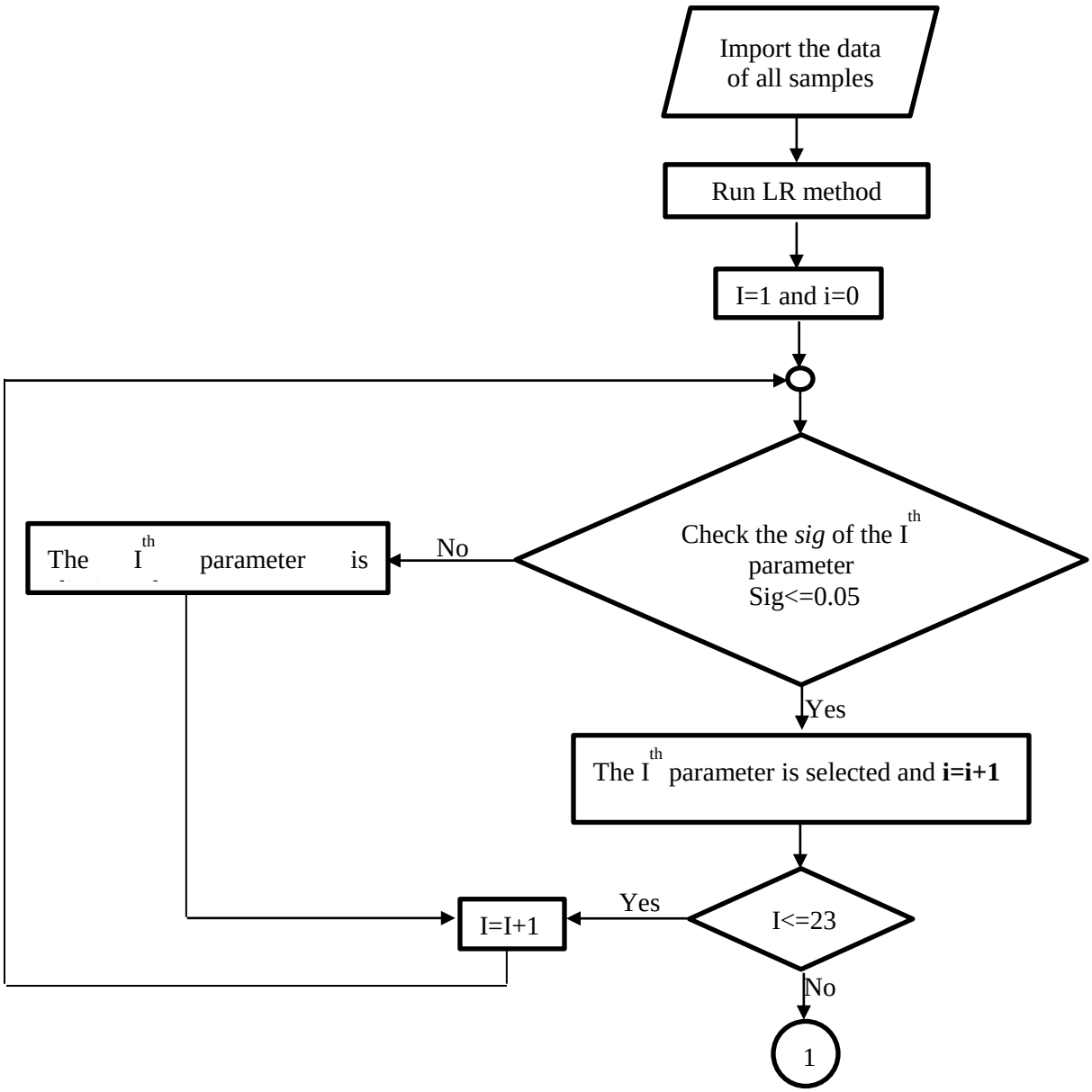
Table 2. List of the risk factors

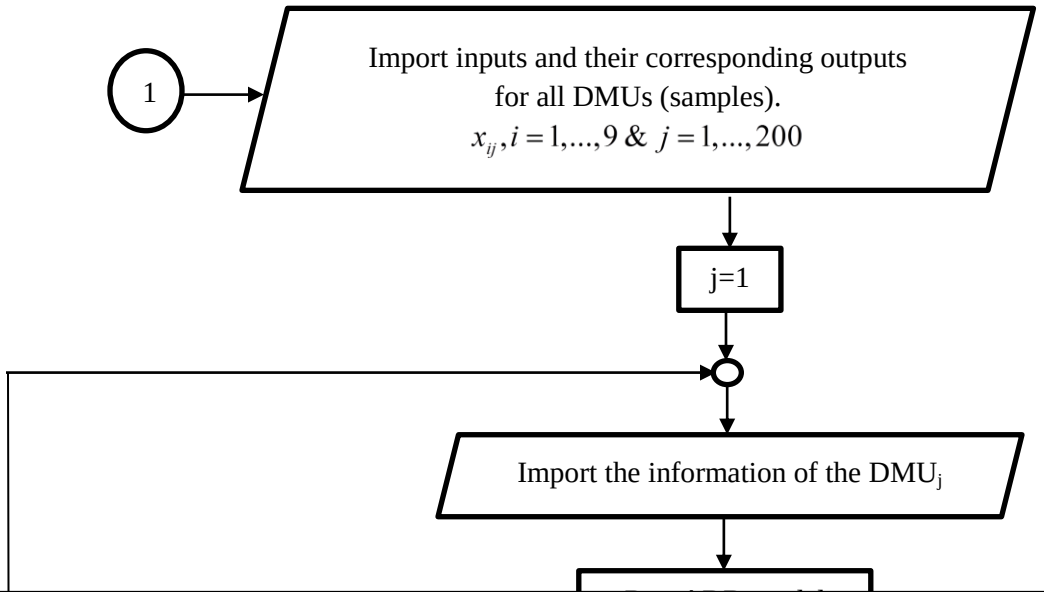
	Name of risk factor		**Name of risk factor**
R_1	Hypertension	R_2	Diabetes Mellitus
R_3	Ischemic Heart Disease	R_4	Atrial Fibrillation
R_5	Heart Valve Disease	R_6	Artificial Heart Valve
R_7	Congestive Heart Failure	R_8	Myocardial Infarction
R_9	Carotid Artery Stenosis	R_{10}	Previous Cerebrovascular Accident
R_{11}	Transient Ischemic Attack	R_{12}	Hyperlipidemia
R_{13}	Vertebrobasilar Insufficiency	R_{14}	Deep Vein Thrombosis
R_{15}	Peripheral Vascular Disease	R_{16}	Head and Neck Trauma
R_{17}	Oral Contraceptive Consumption	R_{18}	Smoking
R_{19}	Other Kinds Of Exposure to Smoke	R_{20}	Addiction
R_{21}	Alcohol Consumption	R_{22}	Snoring
R_{23}	Pregnancy/Delivery/Upto6weeksPostDelivery		

5.3. The Flowchart of the Hybrid Algorithm

The proposed algorithm has been applied to recognize the stroke. For convenience, this approach was shown in a flowchart in Figure 1.







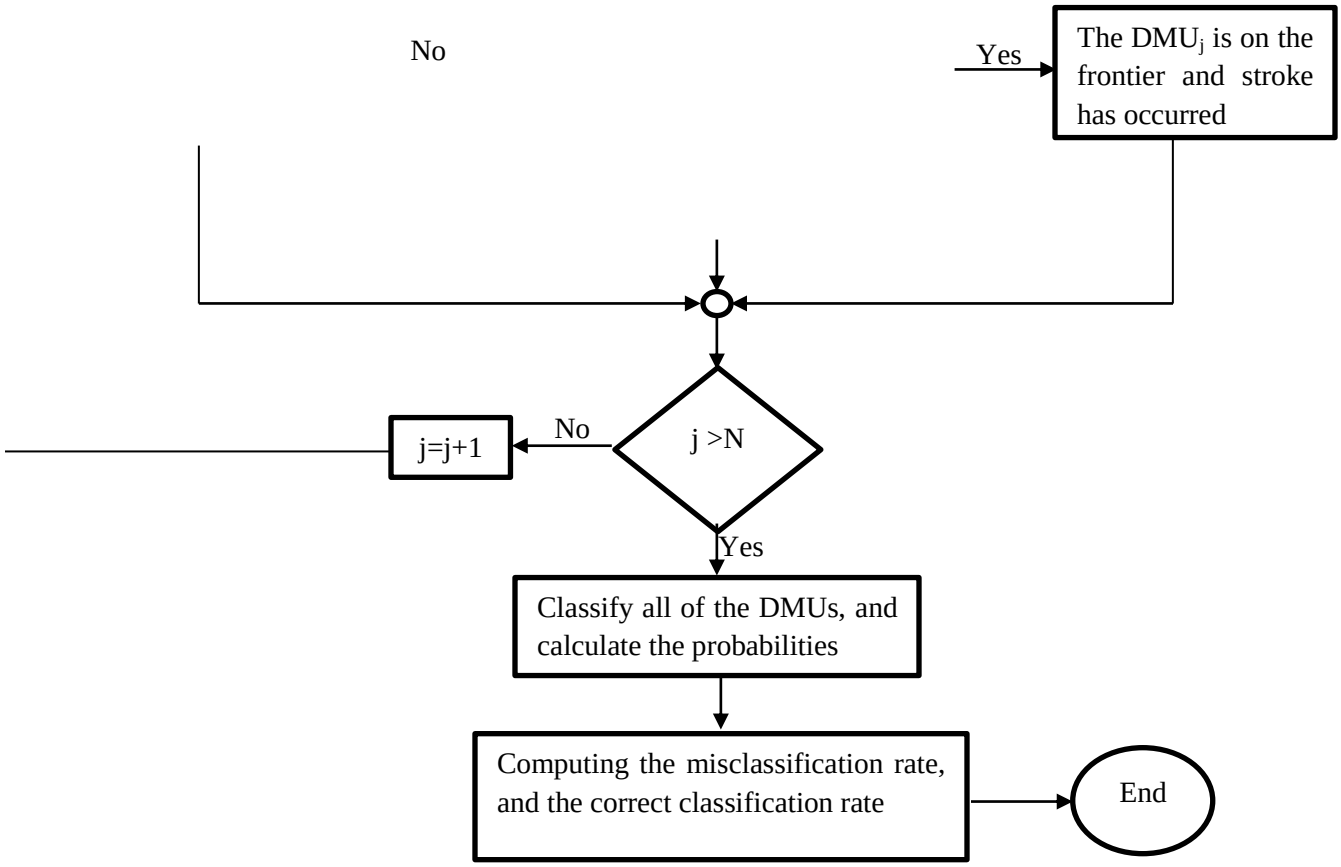


Fig 1. Flowchart of the hybrid algorithm to stroke disease

5.4. Results of the Proposed Approach

We selected 200 patients randomly and studied their cases. It is unfortunate that due to the confidentiality of the information, we are not allowed to publish all the results, and in this paper, only the information of five patients, listed in Table 3, is presented to illustrate the type of actual observations.

Table 3. The information of the five patients

DMUs	x_1	x_2	x_3	x_4	x_5	y_1
P_1	1	0	0	0	1	1

P_2	0	0	0	0	1	1
P_3	1	1	0	0	0	1
P_4	1	0	1	1	0	1
P_5	1	0	1	0	1	0

The results obtained from the implementation of steps 1 to 5 of the algorithm are listed in Table (4).

Table 4. Variables in the equation

Risk factors	B	S.E.	Wald	df	Sig.	Exp(B)
R_1: Hypertension	-.534	.118	20.349	1	.000	.586
R_2: Diabetes Mellitus	-.321	.074	18.873	1	.000	.725
R_4: Atrial Fibrillation	-.759	.219	11.986	1	.001	.468
R_7:Congestive Heart Failure	-1.388	.720	3.721	1	.050	.250
R_{10}:Previous Cerebrovascular Accident	-1.362	.718	3.598	1	.050	.256
R_{12}: Hyperlipidemia	.573	.143	16.037	1	.000	1.773
R_{18}: Smoking	-.276	.125	4.869	1	.027	.759
R_{19}: Other Kinds Of Exposure to Smoke	-.297	.150	3.953	1	.047	.742
R_{22}: Snoring	-.151	.060	6.405	1	.011	.860
Constant	44.807	15886.604	.000	1	.998	2.879E + 19

Columns **B** and **S.E.** show the unstandardized regression coefficients, and standard error, respectively. The **Wald** test ("**Wald**" column) is used to determine statistical significance for each of the independent variables. The statistical significance of the test is found in column "**Sig.**". **Exp(B)** column is the standardized regression coefficient that is used to interpret the results. As can be seen, the information in the "**Variables in the Equation**" table is used to predict the probability of an event occurring based on a one-unit change in an independent variable when all other independent variables are kept constant. According to the WALD test column and the significant level, only 9 factors of 23 factors are effective in the stroke, and the others do not have a direct impact on this disease.

Now, according to step 6 of the algorithm, we should import inputs and their corresponding output based on the parameters that were selected from step 5 for all DMUs. It should be noted that the outputs have two modes: Stroke (=1) or No-Stroke (=0). According to the noted in step 6, to run

ADD model, we first eliminate the data that contains the error of the registry. Consider Table (5) that represents the status of observations after implementing step 7.

Table 5. The status table viewed

Status	One risk factor	More than one risk factor	Stroke with no risk factor (one output)	No stroke with no risk factor (zero output)
Stroke	38	120	20	0
No-Stroke	5	14	0	3

The second and the third columns of the table show the data for the patients who have had a stroke (and No-stroke) by one or more risk factors. As can be seen from Table 5, there are 23 patients (fourth and fifth columns) without risk factors whose information was missing, in which 20 have had a stroke and the other 3 have not.

Now in this stage, we specify DMUs that are on the frontier based on the results extracted from step 7 and using the values of the slacks. Table (6) reports the results of step 8 of the algorithm.

Table 6. Summary of DEA results for Stroke frontier

	The DMUs who place on the frontier (F)	The DMUs who not place on the frontier (NF)	Total
Stroke	$(S \bigcap F) = 158$	$(S \bigcap NF) = 0$	158
No Stroke	$(NS \bigcap F) = 23$	$(NS \bigcap NF) = 19$	42
Total	181	19	200

According to Table (6), those DMUs which are on the frontier have had a stroke by one or more risk factors. This means that, if the DMU has had a stroke by one risk factor and all of the slacks are zero, then it is strongly efficient. In other words, those DMUs which have had a stroke by more risk factors and are on the frontier are weakly efficient, because at least one of their slacks is non-zero. We classify all of the DMUs into four groups. As can be seen from the second column, 158 patients have had a stroke (S) and are on the frontier (F); And 23 patients haven't had a stroke (NS) and are on the frontier (F). The third column shows that there isn't a patient who has had a stroke (S) and isn't on the frontier (NF), and 19 patients haven't had a stroke (NS) and aren't on the frontier (NF).

At the final step, we compute the probabilities related to the four mentioned groups as follows:

(1): $P\left(\frac{S \bigcap F}{S}\right) =$ Divide the number of patients who have had a stroke and are on the frontier by the total number of stroke patients.

(2): $P\left(\frac{S \bigcap NF}{S}\right) =$ Divide the number of patients who have had a stroke and aren't on the frontier by the total number of stroke patients.

(3): $P\left(\frac{NS \bigcap F}{NS}\right) =$ Divide the number of patients who haven't had a stroke and are on the frontier by the total number of no-stroke patients.

(4): $P\left(\frac{NS \cap NF}{NS}\right) =$ Divide the number of patients who haven't had a stroke and aren't on the frontier by the total number of no-stroke patients.

Table (7) shows the results obtained from the probabilities:

Table 7. The probabilities related to the four groups

Formulate the probability	Rate of the probability
$P\left(\frac{S \cap F}{S}\right) = \frac{158}{158}$	**1.00**
$P\left(\frac{S \cap NF}{S}\right) = \frac{0}{158}$	**0.00**
$P\left(\frac{NS \cap F}{NS}\right) = \frac{23}{42}$	**0.5476**
$P\left(\frac{NS \cap NF}{NS}\right) = \frac{19}{42}$	**0.4524**

Table (8) presents the results of the classification of the prediction.

Table 8. Classification table of the prediction

	Predicted		
	Stroke	**No Stroke**	**Percentage Correct**
Stroke	158	0	100%
No Stroke	23	19	45.2%
Overall	(158+19)/200		88.5%

Table 8 shows the information for 158 people who have had a stroke and this result is exactly the same as the real observations. This means that the prediction is done accurately for those who had a stroke. According to the prediction method,19 people have had No-stroke, and this result also is the same as the actual observations. The comparison of the predicted results and real observation for all the under-study patients shows that our proposed model can predict the stroke or no-stroke with an accuracy of 88.5% and this proves that our proposed model is relatively strong. The value for the accuracy can be obtained from the following formula:

$$I_{CC} = \left(\frac{(S \cap F) + (NS \cap NF)}{N}\right) = \left(\frac{158 + 19}{200}\right) = 88.5\%$$

The percentage of the misclassification rate is 11.5% and obtained from the following formula:

$$I_{MC} = \left(\frac{(NS \cap F) + (S \cap NF)}{N}\right) = \left(\frac{23 + 0}{200}\right) = 11.5\%$$

It should be noted that a part of this error can be due to the data registry misinformation.

The results of our study show that the proposed hybrid method has good reliability. It is believed that it can be also used for predicting other similar diseases.

6. Conclusion

In healthcare, prediction is critical to patients and families, as well as professionals and governments. So far, research and studies have been conducted on the application of statistical and analytical methods in health, but less attention has been given to the hybrid methods such as DEA and LR for prediction in health care. This study offered a hybrid algorithm, a combination of ADD model of DEA approach and LR method, and applied it to predict the occurrence or non-occurrence of the stroke. After consulting with the neurologists and reviewing the literature, 23 qualitative risk factors that were thought to be more effective for stroke were selected in this study. 200 samples were used in the hybrid proposed algorithm. Among the 23 risk factors, 9 risk factors were recognized as effective parameters to cause the stroke. The results of this study show that the proposed approach was able to predict the occurrence of the stroke with an accuracy of 88.5%. The proposed method can be used to control, treat, and improve stroke risk factors through timely diagnosis and treatment. This method can also be generalized in various fields.

References

[1] Alinezhad, A. (2016), An integrated DEA and data mining approach for performance assessment, *Iranian Journal of Optimization*, 8 (2), 59-69.

[2] Arasteh, A. (2016), Considering stochastic and combinatorial optimization, *Iranian Journal of Operations Research*, 7(1), 69-84.

[3] Banker, R. D., Charnes, A., and Cooper, W. W. (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, 1078-1092.

[4] Charnes, A., Cooper, W. W., and Rhodes, E. (1978), Measuring the efficiencies of DMUs, *European Journal of Operations Research*, 429- 444.

[5] Charnes, A., Cooper, W. W., Golany, B., and et al. (1985), Foundations of data envelopment analysis for Pareto Koopmans efficient empirical production functions, *Journal of Econometrics*, .30, 91-107.

[6] Dharani, N.P., Bojja, P., and Kumari, P. R. (2021), Evaluation of the performance of an LR and SVR models to predict COVID-19 pandemic, *Journal of Materials Today: Proceedings*, DOI: https://doi.org/10.1016/j.matpr.2021.02.166.

[7] Deljavan, R., Farhoudi, M., and Sadeghi-Bazargani, H. (2018), Stroke in-hospital survival and its predictors: the first results from Tabriz Stroke Registry of Iran, *International Journal of General Medicine*, 11, 233—240.

[8] Freeman, R. V., Eagle, K. A., Bates, E. R., and et al. (2000), Comparison of artificial neural networks with logistic regression in the prediction of in-hospital death after percutaneous transluminal coronary angioplasty, *American Heart Journal*, 140 (3), 511-520.

[9] Fukunishi, H., Nishiyama, M., Luo, Y., and et al. (2020), Alzheimer-type dementia prediction by sparse logistic regression using claim data, *Computer Methods and Programs in Biomedicine*, 196, 105582.

[10] Horváthová, J., and Mokrišová, M. (2020), Comparison of the results of a data envelopment analysis model and logit model in assessing business financial health,

Information, *11*(3), 160, DOI: https://doi.org/10.3390/info11030160

[11] Izadi, B., Ranjbarian, B., Ketabi, S., and Nassiri-Mofakham, F. (2013), Multi-group classification using interval linear programming, *Iranian Journal of Operations Research*, 4(1), 55-74.

[12] Jee, S. H., Park, J. W., Lee, S. Y., and et al. (2008), Stroke risk prediction model: A risk profile from the Korean study, *Atherosclerosis.* 197, 318–325.

[13] Khalili Araghi, M. (2012), Evaluating predictive power of data envelopment analysis technique compared with logit and probit models in predicting corporate bankruptcy, *Australian Journal of Business and Management Research*, 2(09), 38-46.

[14] Karamali, L., Memariani, A., and Jahanshahloo, G. R. (2013), ANN-DEA integrated approach for sensitivity analysis inefficiency models, *Iranian Journal of Operations Research*, 4(1), 14-24.

[15] Khanam, J. J., and Y.Foo, S. (2021), A comparison of machine learning algorithms for diabetes prediction, ICT Express, DOI: https://doi.org/10.1016/j.icte.2021.02.004.

[16] Liew, P. L., Lee, Y. C., Lin, Y. C., and et al. (2007), Comparison of artificial neural networks with logistic regression in the prediction of gallbladder disease among obese patients, *Digestive and Liver Disease*, 39 (4), 356-362.

[17] Lattanzi, S., Pulcini, A., Corradetti, T., and et al. (2020), Prediction of outcome in embolic strokes of undetermined source. J*ournal of Stroke and Cerebrovascular Diseases*, 29 (1), 104486.

[18] Mauthe, R. W., Haaf, D. C., Hayn, P., and et al. (1996), Predicting discharge destination of stroke patients using a mathematical model based on six items from the functional independence measure, *Academy of Physical Medicine and Rehabilitation*. 77, 100-103.

[19] Mendelova, V., and Stachova. M. (2016), Comparing DEA and logistic regression in corporate financial distress prediction, *International Scientific Conference FERNSTAT*.

[20] Mousavi, M. M., Ouenniche, J., and Tone, K. (2019), A comparative analysis of two-stage distress prediction models, *Expert Systems With Applications*, 119, 322- 341.

[21] Nguyen, T., Malley, R., Inkelis, S. H., and Kuppermann, N. (2002), Comparison of prediction models for adverse outcome in pediatric meningococcal disease using artificial neural network and logistic regression analyses, *Journal of Clinical Epidemiology*, 55 (7), 687-695.

[22] NouriJelyani, K., Mohammad, K., Azam, K., and et al. (2011), Application of mixed logistic regression model in determination of effective factors related to visible goiter with the health survey data, *Journal of North Khorasan University of Medical Sciences*, 13, 191- 200.

[23] Nusinovici, S., Tham, Y. C., Yan, M. Y., and et al. (2020), Logistic regression was as good as machine learning for predicting major chronic diseases, *Journal of Clinical Epidemiology*, 122, 56-69.

[24] Premachandra, I. M., Bhabra, G. S., and Sueyoshi, T. (2009), DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique, *European Journal of Operational Research*, 193, 412–424.

[25] Premachandra, I. M., and Yao Chen, J. (2011), DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment, *Omega*, 32, 620–626.

[26] Pourmahmoud, J., and GholamAzad, M. (2021), Data envelopment analysis using the binary-data, *Journal of Modelling in Management*, DOI: https://doi.org/10.1108/JM2-10-2019-0246.

[27] Ruczinski, I., and Kooperbery, C. (2003), Logistic Regression, *Journal of Computational and Graphical Statistics*, 12, 475-511.

[28] Reed, P., and Wu, Y. (2013), Logistic regression for risk factor modeling in stuttering research, *Journal of Fluency Disorders*, 38 (2), 88-101.

[29] Radovanovic, S., Radojicic, M., and Savic, G. (2014), Two-phased DEA-MLA approach for predicting the efficiency of NBA players, *Yugoslav Journal of Operations Research*, 24 (3), 347-358

[30] Sil Va E Souza, G., and Goncalves Gomes, E. (2014), Assessing the significance of covariates in output-oriented data envelopment analysis with two-stage regression models, *Wests Transactions Ones on Systems*, 13, 440-449.

[31] Sergeev, A. V., and Weckman, G. R. (2015), Cardiovascular disease treatment outcomes in patients with diabetes: Prediction models using artificial neural networks and logistic regression, *Annals of Epidemiology*, 25(9), 705.

[32] Shukla, Sh., Jain, R., and Dubey, Sh. (2017), Med alert- A diabetic predictor using logistic regression statistical model, *International journal for research in applied science and engineering technology*, 5, 332-334.

[33] Sposato, L. A., Stirling, D., and Saposnik, G. (2018), Therapeutic decisions in atrial fibrillation for stroke prevention: The pole of aversion to ambiguity and physicians' risk preferences, *J Stroke Cerebrovasc*, 27(8),2088-2095.

[34] Zhu, C., Idemudia, C. U., and Feng. W. (2019), Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, *Informatics in Medicine Unlocked*, 17, 100179.

[35] Zhu, N., Zhu, Ch., and Emrouznejad, A. (2020), A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies, *Journal of Management Science and Engineering*, DOI: https://doi.org/10.1016/j.jmse.2020.10.001