Egyptian Informatics Journal 22 (2021) 145–153

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

The fuzzy common vulnerability scoring system (F-CVSS) based on a least squares approach with fuzzy logistic regression

Kerem Gencer [a,*], Fatih Başçiftçi [b]

[a] Karamanoglu Mehmetbey University, Vocational School of Technical Sciences, Department of Computer Programming, Karaman, Turkey
[b] Selcuk University, Faculty of Technology, Department of Computer Engineering, Konya, Turkey

ARTICLE INFO

Article history:
Received 1 April 2020
Revised 29 June 2020
Accepted 8 July 2020
Available online 28 July 2020

Keywords:
Fuzzy logistic regression
Probability rates
Fuzzy least squares
Common Vulnerability Scoring System

ABSTRACT

This study presents a new approach for calculations within the Common Vulnerability Scoring System that scoring the effects of vulnerabilities in software on the security status. These calculations is the method that is most commonly used in scoring software vulnerabilities. The present model demonstrates how software security vulnerabilities can be calculated using linguistic terms. Therefore, the proposed method has a more flexible structure than this system. The current Common Vulnerability Scoring System formula and scores were used to assess and implement the presented model. The aim was to form a fuzzy model called the Fuzzy Common Vulnerability Scoring System based on the success probabilities which are defined using linguistic terms such as low, very low or high. Moreover, the Fuzzy Logistic Regression (FLR) method was used to define the relationship between the exact inputs and fuzzy multiple outputs, and the Least Squares Method was used to estimate the parameters of the presented model. The performance of the model was evaluated by a comparison using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Kim and Bishu's criterion. Validity of the fuzzy regression model is demonstrated with different fitness functions. The expectation was that more practical estimations with better error tolerance can be achieved by using linguistic terms to assess common vulnerabilities.

1. Introduction

Vulnerability is defined as an error or event that may directly or indirectly distort the integrity, practicality or privacy in a system, device or service. Many vulnerabilities are detected and declared every year. The administrators of organizations with the resources to terminate vulnerabilities in their systems first focus on the most critical vulnerabilities. The number of companies and systems assessing these risks is quite limited. The Common Vulnerability Scoring System (CVSS), based on the Forum of Incident Response and Security Teams (FIRST) open platform, is a quantitative calculation method transforming simple formulae into a value of the severity of vulnerabilities [1]. The availability of standard metrics allows an objective assessment to be performed to compare vulnerabilities. Therefore, the use of CVSS is very beneficial. CVSS has three metrics groups: basic, temporary and environmental. However, vulnerability databases are generally associated with base points. The basic score represents the characteristics that are not affected by the concepts of place, system and time, and that are present in the nature of all vulnerabilities. Base metric consists of 5 basic components. Attack Vector reflects the context by which vulnerability exploitation is possible. Attack Complexity describes the conditions beyond the attacker's control that must exist in order to exploit the vulnerability. Privileges Required describes the level of privileges an attacker must possess before successfully exploiting the vulnerability. User Interaction captures the requirement for a user, other than the attacker, to participate in the successful compromise of the vulnerable component. The Impact metrics refer to the properties of the impacted component. If a scope change has not occurred, the Impact metrics should reflect the confidentiality, integrity, and availability (CIA) impact to the vulnerable component [2]. Fig. 1 indicates the basic CVSS formula. Despite the fact that CVSS is a standard metrics system, there are a couple of limitations. CVSS calculation includes subjective calculations as indicated in Fig. 1. Moreover, users need to be competent in terms of both the vulnerability characteristics and the CVSS

* Corresponding author.
E-mail addresses: keremgen@kmu.edu.tr (K. Gencer), basciftci@selcuk.edu.tr (F. Başçiftçi).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

https://doi.org/10.1016/j.eij.2020.07.001
1110-8665/

scoring systems. Explaining vulnerability deficits through linguistic variables is easier than finding and calculating the basic CVSS scoring parameters [3]. This study aimed to model and estimate CVSS base scores based on a vulnerability description using Fuzzy Logistic Regression (FLR).

Regression analysis is one of the popular methods used to model the relationship between a dependent variable (output) and one – or more than one – independent variable (input/inputs). Where the meaning of certain terms may cause complication or uncertainty in real life, these terms should be defined using linguistic statements or fuzzy numbers. Linguistic variables turned into the means requiring the use of fuzzy sets to express verbal statements digitally. Ref. [4] offered the fuzzy regression model where certain terms are represented by fuzzy numbers Fuzzy regression is based on the theory of fuzzy sets and fuzzy probability. In published literature, it has been commonly used in two forms. The first is based on the concepts of probability suggested by Tanaka, and it reduces the total uncertainty of estimated values for the outputs [5–7]. In contrast, the second uses the approach of minimizing errors between the simultaneously observed outputs, which is the fuzzy Least Squares Method suggested by Diamond and Celmiņš [8–11].

This study proposed the use of the FLR model along with the exact input and fuzzy output. Diamond's FLS approach was used to estimate the parameters of the fuzzy logistic regression model. Section 2 includes some studies from relevant literature. Certain definitions regarding fuzzy sets and a brief explanation of the classic concept of logistic regression are included as a preamble to the suggested method in Section 3. The FLR methods are reviewed in Section 4, and the FLR model was formed using the CVSS dataset in Section 5. Criterias were compared to test the suitability of the suggested FLR model in Section 6, and the results are presented in Section 7.

2. Literature review

The number of studies explaining vulnerabilities through linguistic variables (fuzziness) is quite limited. Ref. [12] presented the fuzzy systems approach to assess the relative potential risk of computer networks being exposed to attacks arising from vulnerabilities. He used this approach to sort the vulnerabilities. Accordingly, the purpose was to get analysts to prioritize their studies based on the potential risk exposure of assets and networks. To assess the vulnerability in telecommunication networks, Ref. [13] suggested an approach based on CVSS V3, expert decisions and fuzzy logic. Ref. [14] presented a holomorphic risk level estimation model as a conditional probability of the frequency and impact estimations regarding CVSS. Logistic regression analysis was used to model one – or more than one – independent (explanatory) variable(s) through a multiple response variable. FLR was used to define the relationship between exact or fuzzy inputs and fuzzy dual/multiple outputs. The number of studies where FLR is used is quite limited.

Diamond's FLS Method and Tanaka's probability approach were used to estimate the parameters of the FLR model used in the present study. FLR was suggested by Pourahmad et al. for dual outputs and exact inputs [15]. Ref. [4] used the concept of a probabilistic method to estimate the parameters of the model and proposed a probability approach and certain suitability criteria. FLR was suggested for modeling the fuzzy dual outputs. Moreover, the logarithmic conversion of probability rates was modeled using the enlargement principle, and Diamond's method was used to estimate the parameters [15]. A new technology credit scoring model based on the FLR method was proposed using fuzzy inputs and fuzzy dual outputs [16]. Exact inputs and fuzzy outputs were used by many authors from published literature for the fuzzy regression models [5,17–19]. Some of the studies where fuzzy outputs, fuzzy inputs and exact parameters were used, were reviewed by [8,10,20] Fuzzy outputs, fuzzy inputs and fuzzy parameters were examined by [11,21]. Recently, [22] propose a novel fuzzy regression method in order to predict affective quality and estimate fuzziness in human assessment, when objective features are given. Ref. [23] propose a systematic approach using the FLR models attached with optimized h values to identify the functional relationships in QFD, where the coefficients are assumed as symmetric triangular fuzzy numbers. Ref. [24] propose a novel chaos-based fuzzy regression (FR) approach with

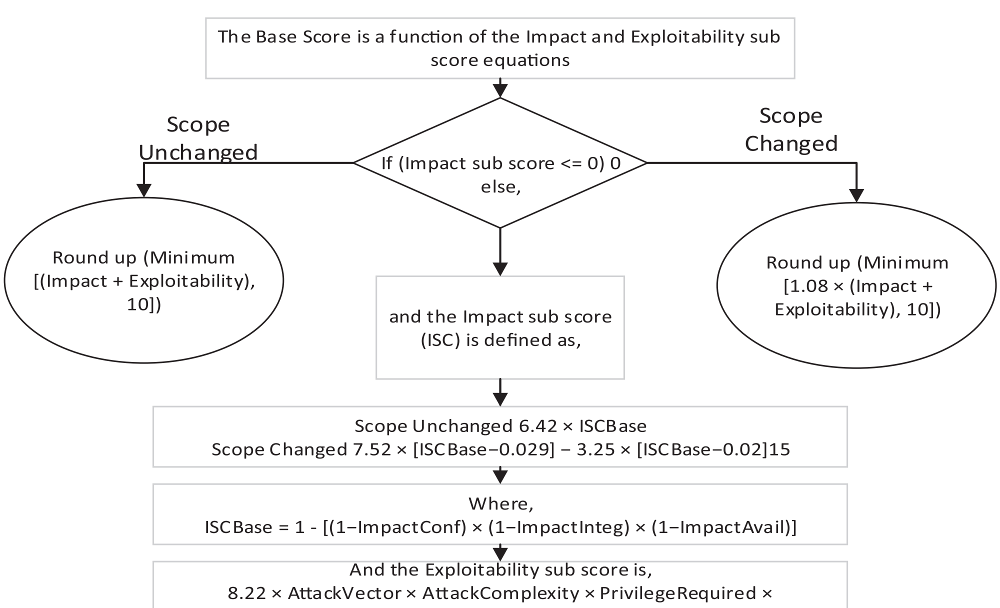

Fig. 1. The CVSS basic scoring formula [2].

which fuzzy customer satisfaction models with second- and/or higher order terms, and interaction terms can be developed. Ref. [25] compare the widely used methodology of discarding some implied volatilities and interpolating the remaining knots with cubic splines, to a fuzzy regression approach which does not require an a-priori choice of implied volatilities. Ref. [26] propose a fuzzy regression method that accounts for fuzziness introduced through human judgment and the limitations of widely-used psychometric quality scales.

3. The underlying preliminary theory for the proposed model

This section provides a short brief on certain definitions regarding arithmetic operations on fuzzy sets and introduces the classic logistic regression method.

3.1. Some basic rules in fuzzy set theory

Definition 3.1.1. *Consider U elements as a universal set indicated by "X". Membership of X, a classic subset of U, is indicated through the μ_A function and with $\widetilde{A}$ fuzzy sets and a regular dual set as follows*:

$$\widetilde{A} = \left\{ \left(x, \mu_{\widetilde{A}}(x) \right) / x \in X \right\} \tag{1}$$

Definition 3.1.2. *α level sets of a fuzzy set defined in X space $\widetilde{A}$ are expressed as follows*:

$$\mu_{\widetilde{A}}(x) \geq \alpha, \ \alpha \in [0, 1] \tag{2}$$

Definition 3.1.3. *A triangular fuzzy figure indicated by the $\widetilde{A} = (a_1, a_2, a_3)$ fuzzy set is mathematically illustrated as follows*:

$$\mu_{\widetilde{A}}(x) = \mu_{\widetilde{A}}(x; a, b, c) = \left\{ \begin{array}{ll} \frac{x-a}{b-a} & if \ \ a \leq x \leq b \\ \frac{c-x}{c-b} & if \ \ b \leq x \leq c \\ 0 & if \ \ x > c \, or \, x < a \end{array} \right\} \tag{3}$$

Definition 3.1.4. *$\widetilde{A} = (a_1, a_2, a_3)$ and $\widetilde{B} = (b_1, b_2, b_3)$ are two triangular fuzzy figures, and some arithmetical operations regarding fuzzy figures are as follows*:

$$\text{Addition/Subtraction:} \widetilde{A} + \widetilde{B} = (a_1 + b_1 / a_1 - b_3, a_2 + b_2 / a_2 - b_2, a_3 + b_3 / a_3 - b_1) \tag{4}$$

$$\text{Multiplication} : \widetilde{A} \times \widetilde{B} = (a_1.b_1, a_2.b_2, a_3.b_3) \tag{5}$$

$$\text{Division} : \widetilde{A} \times \widetilde{B} = (a_1/b_1, a_2/b_2, a_3/b_3) \tag{6}$$

Definition 3.1.5. *Let F indicates fuzzy sets space (therefore, $E \subseteq F$. So, for each arbitrary $m \in F, m : R \to [0, 1]$, reflects X to be Cartesian product of universes $(X_1, X_2, \cdots, X_n)$i.e. $(X_1 \times \cdots \times X_n)$ and $(m_1, m_2, \cdots, m_n)$ are n fuzzy sets in $X_1, X_2, \cdots, X_n$ respecticely. Also, suppose that f is a mapping from X to a universe Y and $y = f(x_1, x_2, \cdots, x_n)$. Then the extension principle lets us to define a fuzzy set in Y.*

$$\widetilde{A} = \left\{ \begin{array}{ll} \sup min\{m_1(x_1), m_2(x_2), \cdots, m_n(x_n)\}, & f^{-1}(y) \neq 0 \\ 0 & , \ otherwise \end{array} \right\} \tag{7}$$

Definition 3.1.6. *Use of approximate figures by decision makers working on sensitive topics is an undesirable approach. Therefore, a subtraction should be made from the new fuzzy set achieved as a result of the operations performed with the fuzzy numbers. In other words, converting the fuzzy output set into an exact value after the operations have been performed with fuzzy figures is called purification. Purification occurs through the membership functions of fuzzy sets. In this study, purification was made using the$y_{ic} = \frac{(y_{il} + y_{im} + y_{ir})}{3}$ equation with the Centroid method, where y_{im} is central, y_{il} is left deviation, and y_{ir} is right deviation* [27].

3.2. Ordinary logistic regression

The purpose in using logistic regression analysis is to form an acceptable model that can define the association between the dual dependent variables and one – or more than one – independent variable(s) in a way that the least number of variables is used and the best adaptation is ensured. Logistic regression is used mainly in medicine, but it is also employed in credit scoring and genetics. No assumptions regarding the distribution of the variables used are present. The response variable may be dual or multiple. Independent variables may be continuous, intermittent, dual or a combination of these. There is no linear relationship between the dependent (output) and independent variables. The success probability is represented by (π) while failure is 0 $(1 - \pi)$ for the dual-category dependent variable. The logit conversion of π is defined as follows:

$$\pi = \frac{1}{1 + e^{b_{0+} b_1 x_1 + \cdots + b_n x_n}} \tag{8}$$

The logarithm of the model in Eq. (8) is as follows:

$$ln\left(\frac{\pi}{1 - \pi}\right) = b_{0+} b_1 x_1 + \cdots + b_n x_n \tag{9}$$

The equation $\frac{\pi}{1-\pi} = e^{b_{0+} b_1 x_1 + \cdots + b_n x_n}$ here is called the probability rate, and b, $x_{ij,}$ $i = 1, 2, \cdots, m$, j = 0,1,2,. . .,n reflects the parameters of the model.

4. Fuzzy logistic regression

Fuzzy logistic regression defines the relationship between two or more dependent variables through exact or fuzzy independent variables. A probabilistic and Least Square approach was used to estimate the parameters in the fuzzy logistic regression model. Published literature contains many studies about fuzzy linear regression, but the research on FLR is quite limited [28]. A model is established in logistic regression, and uncertain or fuzzy observations are terminated. However, the model cannot be established if there are too many fuzzy observations. Therefore, the concept of probability rates is used for these observations.

4.1. The proposed model

If $\widetilde{Y}_i, i = 1, 2, \cdots, m$ fuzzy multiple dependent variables and $x_{ij,}$ $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, n$ are regarded as a set of exact independent variables, the definition of probability rates is as follows:

$\mu_i = i = 1, 2, \cdots, m$ are the success probabilities, while $\left(\mu_i = poss\widetilde{Y}_i = 1 \right)$ and the aforenoted equation are exact figures. They are expressed as $\mu \in R, 0 \leq \mu \leq 1$,or through the linguistic thresholds (very low, low, moderate, high, very high).

Based on the $\frac{\widetilde{\mu}_i}{1 - \widetilde{\mu}_i}, i = 1, 2, \cdots, m$ equation, i is defined as the probability rate for observation.

In this method, as a response to $\tilde{\mu}_i$, $\tilde{Y}_i$ triangular fuzzy numbers are calculated with the formula $\tilde{Y}_i = \frac{\tilde{\mu}_i}{1-\tilde{\mu}_i}$. The fuzzy multiple dependent variable model is defined as follows:

$$\tilde{W}_i = ln\left(\frac{\tilde{\mu}_i}{1-\tilde{\mu}_i}\right) = A_0 + A_1x_{i1} + \cdots + A_nx_{in}i = 1,2,\cdots,m \quad (10)$$

We suppose $A_i = (a_i s_i)_{Triangular}$ $i = 1,2,\cdots,m$ in Eq. (10). Then the estimates outputs are symmetric triangular fuzzy numbers $\tilde{W}_i = (f_i(\alpha), f_i(s))_T$. $\tilde{W}_i$ is the estimated logarithmic conversion for the probability rates and is called the fuzzy dependent variable.

where $f_i(\alpha) = a_0 + a_1x_{i1} + \cdots + a_nx_{in}$ and (11)

$$f_i(s) = s_0 + s_1x_{i1} + \cdots + s_nx_{in} \quad (12)$$

$\tilde{W}_i$ is triangular fuzzy numbers and is denoted by $(f_i(\alpha), f_i(s))_T$ symbolically [15] are expressed as follows:

$$\tilde{W}_i = (f_i(\alpha), f_i(s))_T, i = 1,2,\cdots,m \quad (13)$$

α level sets of $\tilde{W}_i$ are calculated as follows:

$$\left(\tilde{W}_i\right)_\alpha = [f_i(\alpha) - (1-\alpha)f_i(s), f_i(\alpha) + (1-\alpha)f_i(s)] \quad (14)$$

$\tilde{W}_i$ is estimated fuzzy logistic model. In this model, the expert person assesses the term $\tilde{\mu}_i$ as success probabilities in this method. The logarithmic conversions of probability rates are regarded as the observed dependent variables. The expression for $\tilde{w}_i$ is as follows:

$$\tilde{w}_i = ln\left(\frac{\tilde{\mu}_i}{1-\tilde{\mu}_i}\right), i = 1,2,\cdots,m \quad (15)$$

Membership functions of $\tilde{w}_i$'s (observed dependent variables or observed outputs) are calculated based on the expansion principle. The α level sets of $\tilde{w}_i$ based on $\tilde{\mu}$ are shown as follows:

$$\left(\tilde{w}_i\right)_\alpha = \left[ln\frac{l_{\tilde{\mu}_i}(\alpha)}{1-l_{\tilde{\mu}_i}(\alpha)}, ln\frac{r_{\tilde{\mu}_i}(\alpha)}{1-r_{\tilde{\mu}_i}(\alpha)}\right] \quad (16)$$

by using following calculations:

$$\left(\tilde{\mu}_i\right)_\alpha = \left[l_{\tilde{\mu}_i}(\alpha), r_{\tilde{\mu}_i}(\alpha)\right] \quad (17)$$

$$l_{\tilde{\mu}_i}(\alpha) = m_{\mu_i} - (1-\alpha)(m_{\mu_i} - l_{\mu_i}) \quad (18)$$

$$r_{\tilde{\mu}_i}(\alpha) = m_{\mu_i} + (1-\alpha)(r_{\mu_i} - m_{\mu_i}) \quad (19)$$

4.2. Estimation of parameters (Table 3)

To attained the Eq. (10) optimally, the Sum of Squared Errors (SSE) between $\tilde{w}_i$ and $\tilde{W}_i$, $\tilde{W}_i, i = 1,2,\cdots,m$, was minimized using Diamond's Least Squares.

$$SSE = \sum_{i=1}^{m} d\left(\tilde{w}_i, \tilde{W}_i\right)^2 = \sum_{i=1}^{m}\int_0^1 \left[f(\alpha)d^2\left(\left(\tilde{w}_i\right)_\alpha, \left(\tilde{W}_i\right)_\alpha\right)d\alpha\right]^{1/2} \quad (20)$$

SSEs are calculated using Eqs. (14) and (16) as follows:

$$SSE = \sum_{i=1}^{m}\int_0^1 f(\alpha)\left[\begin{array}{l}\left[ln\frac{l_{\tilde{\mu}_i}(\alpha)}{1-l_{\tilde{\mu}_i}(\alpha)} - f_i(\alpha) - (\alpha-1)f_i(s)\right]+ \\ \left[ln\frac{r_{\tilde{\mu}_i}(\alpha)}{1-r_{\tilde{\mu}_i}(\alpha)} - f_i(\alpha) - (1-\alpha)f_i(s)\right]\end{array}\right]^2 d\alpha \quad (21)$$

The total for a_j and s_j error is minimized for $j = 1,2,\cdots,n$ and is calculated using the equations below:

$$\frac{\partial SSE}{\partial a_j} = 0, \frac{\partial SSE}{\partial s_j} = 0 \quad (22)$$

The leads to the following equations.

$$\frac{\partial SSE}{\partial a_j} = \sum_{i=1}^{m}\left(\int_0^1 2\alpha x_{ij}\left[2f_i(a) - ln\frac{l_{\tilde{\mu}_i}(\alpha)}{1-l_{\tilde{\mu}_i}(\alpha)} - ln\frac{r_{\tilde{\mu}_i}(\alpha)}{1-r_{\tilde{\mu}_i}(\alpha)}\right]d\alpha\right)$$

$$\frac{\partial SSE}{\partial s_j} = \sum_{i=1}^{m}\left(\int_0^1 2\alpha(1-\alpha)x_{ij}\left[2(1-\alpha)f_i(s) + ln\frac{l_{\tilde{\mu}_i}(\alpha)}{1-l_{\tilde{\mu}_i}(\alpha)} - ln\frac{r_{\tilde{\mu}_i}(\alpha)}{1-r_{\tilde{\mu}_i}(\alpha)}\right]d\alpha\right) \quad (23)$$

Success probabilities of $\tilde{\mu}_i$ based on Eqs. (20) and (21) are assigned by an expert for every observation. The following equation is achieved as a result of these assessments and related integrals:

$$a_0\sum_{i=1}^{m} x_{i0}x_{ij} + a_1\sum_{i=1}^{m} x_{i1}x_{ij} + \cdots + a_n\sum_{i=1}^{m} x_{in}x_{ij} = \sum_{i=1}^{m} z_ix_{ij}, j = 1,2,\cdots,n$$

$$s_0\sum_{i=1}^{m} x_{i0}x_{ij} + s_1\sum_{i=1}^{m} x_{i1}x_{ij} + \cdots + s_n\sum_{i=1}^{m} x_{in}x_{ij} = \sum_{i=1}^{m} k_ix_{ij}, j = 1,2,\cdots,n \quad (24)$$

$x_{i0} = 1$, $i = 1,2,\cdots,m$, and the terms z_i and k_i are a result of integral calculations for every observation. Eq. (23) a is reflected in the matrix form as follows:

Based on $aA = Z$ and$sA = K$, in which

$$A = X'X, X = \begin{bmatrix} 1 & x_{11} & \cdots. & x_{1n} \\ 1 & x_{21} & \cdots. & x_{2n} \\ \cdots. & \cdots. & \cdots. & \cdots. \\ \cdots. & \cdots. & \cdots. & \cdots. \\ 1 & x_{m1} & \cdots. & x_{mn} \end{bmatrix}_{m\times(n+1)} \quad (25)$$

$$a = (a_0, a_1, \cdots, a_n)$$
$$Z = \left(\sum_{i=1}^{m} z_ix_{i0}, \sum_{i=1}^{m} z_ix_{i1}, \sum_{i=1}^{m} z_ix_{i2}, \cdots, \sum_{i=1}^{m} z_ix_{in}\right)^T \quad (26)$$

$$s = (s_0, s, \cdots, s_n)$$
$$K = \left(\sum_{i=1}^{m} k_ix_{i0}, \sum_{i=1}^{m} k_ix_{i1}, \sum_{i=1}^{m} k_ix_{i2}, \cdots, \sum_{i=1}^{m} k_ix_{in}\right)^T \quad (27)$$

$$a = A^{-1}Z$$

$$s = A^{-1}K \quad (28)$$

As a result of, the optimal fuzzy logistic model is expressed as follows:

$$W_i = (a_0, s_0) + (a_1, s_1)x_{i1} + \cdots + (a_n, s_n)x_{in}$$
$$i = 1,2,\cdots,m \quad (29)$$

5. Application

The linguistic thresholds are first determined (as very low, low, moderate, high and very high). As a response to $\tilde{\mu}_i$, $\tilde{Y}_i$ triangular fuzzy figures were calculated, and α level sets of $\tilde{W}_i$ were found. Membership functions for $\tilde{w}_i$ were calculated based on the expansion principle. The SSEs between $\tilde{w}_i$ and $\tilde{W}_i$ were minimized using Diamond's Least Squares Method. The fuzzy logistic regression model based on probability rates was later suggested. There are basically seven input variables and one output variable representing the essence of the vulnerability scoring system in Table 1. The output variable has five categories and is indicated in Table 2. Decision makers cannot allude to certainty regarding the score that was exactly calculated, even if a particular calculation approach was adopted. Classic logistic regression cannot be implemented due to the fuzziness of CVSS severity. FLR is suitable in this case.

The inputs are: Attack Vector (AV), Attack Complexity (AC), Privileges Required (PR), User Interaction (UI), Confidentiality (C), Integrity (I), Availability (A), and the 5 categories of outputs are: Critical, High, Medium, Low, Very Low. FLR was implemented using the values in the calculations made with CVSS 3.0 in Table 6. Y: Classification variable. 0 to 0.09 reflects a very low severity while 0.1 to 3.9 means low severity, 4 to 6.9 indicates moderate severity, 7 to 8.9 means high severity and 9 to 10 reflects critical severity.

Based on the degrees of $\tilde{\mu} \in$: Critical, High, Medium, Low and Very Low; fuzzy multi-category, $\tilde{\mu}$, was regarded as a fuzzy dependent observed variable ($\tilde{\mu} \in$ Critical, High, Medium, Low, Very Low). Probability rates were calculated as a fuzzy dependent variable. The triangular fuzzy figures for $\tilde{\mu}$, approximated 'exact' scores and linguistic terms are presented in the Table 3 below.

Table 1
Metric values regarding CVSS.

Inputs	Independent Variables	Coefficients
Attack Vector (AV)	Network	0.85
	Adjacent	0.62
	Network	0.55
	Local	0.20
	Physical	
Attack Complexity (AC)	Low	0.77
	High	0.44
Privileges Required (PR)	None	0.85
	Low	0.62 (0.68 If Scope/Modified Scope is Changed)
	High	0.27 (0.50 If Scope/Modified Scope is Changed)
User Interaction (UI)	None	0.85
	Required	0.62
C, I, A Impact	High	0.56
	Low	0.22
	None	0

Table 2
Severity rating scale regarding CVSS.

Linguistic Terms	Lowest	Highest
Very Low	0	0.09
Low	0.1	3.9
Medium	4	6.9
High	7	8.9
Critical	9	10

Seven exact inputs and five fuzzy outputs were used to form the FLR model. Probability measurements and logarithmic conversion of probability rates for fuzzy outputs were performed (see below).

For i = High ($\tilde{\mu}_{High}$ approximately "High"), the estimated $\tilde{W}_1$ was as follows:

$$\tilde{W}_{High} = ln\left(\frac{\tilde{\mu}_1}{1-\tilde{\mu}_1}\right) = A_0 + A_1x_{11} + \cdots + A_7x_{17} = (f_1(\alpha), f_1(s))_T \tag{30}$$

α level sets regarding $\tilde{W}_{High}$ are:

$$\left(\tilde{W}_{High}\right)_\alpha = [f_1(a) - (1-\alpha)f_1(s), f_1(\alpha) + (1-\alpha)f_1(s)] \tag{31}$$

using the equations below:

$$f_{High}(\alpha) = a_0 + a_1x_{11} + \cdots + a_7x_{17} = a_0 + 0.85a_1 + \cdots + 0a_7$$

$$f_{High}(s) = s_0 + s_1x_{11} + \cdots + s_7x_{17} = s_0 + 0.85s_1 + \cdots + 0s_7 \tag{32}$$

The α level sets for $\tilde{w}_{High}$ were calculated as follows:

$$\left(\tilde{w}_{High}\right)_\alpha = \left[ln\frac{l_{\tilde{\mu}_1}(\alpha)}{1-l_{\tilde{\mu}_1}(\alpha)}, ln\frac{r_{\tilde{\mu}_1}(\alpha)}{1-r_{\tilde{\mu}_1}(\alpha)}\right] \tag{33}$$

using the equation below:

$$\left(\tilde{\mu}_{High}\right) = \left[l_{\tilde{\mu}_1}(\alpha), r_{\tilde{\mu}_1}(\alpha)\right] = [0.795 - (1-\alpha)0.095, 0.795 + (1-\alpha)0.095] \tag{34}$$

SSEs regarding $\tilde{w}_{High}$ and $\tilde{W}_{High}$ are shown by:

$$SSE_{High} = \int_0^1 f(\alpha) + \left[ln\frac{0.795-(1-\alpha)0.095}{0.205+(1-\alpha)0.095} - f_1(\alpha) - (\alpha-1)f_1(s)\right]^2 + \left[ln\frac{0.795+(1-\alpha)0.095}{0.205-(1-\alpha)0.095} - f_1(\alpha) - (\alpha-1)f_1(s)\right]^2 d\alpha \tag{35}$$

For i = Medium ($\tilde{\mu}_{Medium}$ approximately "Medium"), the estimated $\tilde{W}_1$ is as follows:

$$\tilde{W}_{Medium} = ln\left(\frac{\tilde{\mu}_1}{1-\tilde{\mu}_1}\right) = A_0 + A_1x_{11} + \cdots + A_7x_{17} = (f_1(\alpha), f_1(s))_T \tag{36}$$

α level sets for $\tilde{W}_{Medium}$:

$$\left(\tilde{W}_{Medium}\right)_\alpha = [f_1(a) - (1-\alpha)f_1(s), f_1(\alpha) + (1-\alpha)f_1(s)] \tag{37}$$

are found by using the equations below:

$$f_{Medium}(\alpha) = a_0 + a_1x_{11} + \cdots + a_7x_{17} = a_0 + 0.55a_1 + \cdots + 0.22a_7$$

$$f_{Medium}(s) = s_0 + s_1x_{11} + \cdots + s_7x_{17} = s_0 + 0.55s_1 + \cdots + 0.22s_7 \tag{38}$$

The α level sets for $\tilde{w}_{Medium}$ are calculated as follows:

$$\left(\tilde{w}_{Medium}\right)_\alpha = \left[ln\frac{l_{\tilde{\mu}_1}(\alpha)}{1-l_{\tilde{\mu}_1}(\alpha)}, ln\frac{r_{\tilde{\mu}_1}(\alpha)}{1-r_{\tilde{\mu}_1}(\alpha)}\right] \tag{39}$$

Table 3
Transformation of triangular fuzzy numbers ($\tilde{\mu}$) and the corresponding $\tilde{Y}$

Linguistic Scales	Triangular Fuzzy Numbers			$\tilde{Y} = ln\left(\frac{\mu}{1-\mu}\right)$		
	Left	Center	Right			
Very Low	0.01	0.05	0.09	−4.5951	−2.9444	−2.3136
Low	0.10	0.245	0.39	−2.1972	−1.1255	−0.4473
Medium	0.40	0.545	0.69	−0.4055	0.1805	0.8001
High	0.70	0.795	0.89	0.8473	1.3553	2.0907
Critical	0.90	0.945	0.99	2.1972	2.8439	4.5951

using the following equation:

$$\left(\tilde{\mu}_{Medium}\right) = \left[l_{\tilde{\mu}_1}(\alpha), r_{\tilde{\mu}_1}(\alpha)\right] = [0.545 - (1-\alpha)0.145, 0.545 + (1-\alpha)0.145] \quad (40)$$

Squared errors regarding $\tilde{w}_{Medium}$ and $\tilde{W}_{Medium}$ are calculated as follows:

$$SSE_{Medium} = \int_0^1 f(\alpha) + \left[ln\frac{0.545-(1-\alpha)0.145}{0.455+(1-\alpha)0.145} - f_1(\alpha) - (\alpha-1)f_1(s)\right]^2 + \left[ln\frac{0.545+(1-\alpha)0.145}{0.455-(1-\alpha)0.145} - f_1(\alpha) - (\alpha-1)f_1(s)\right]^2 d\alpha \quad (41)$$

After the SSEs were obtained for all observations, a_j and s_j's were calculated using Eq. (23). The terms a and s were found as follows:

$$a = A^{-1}Z = (-3.9623, 1.0909, 1.0519, 1.0964, 0.9262, 1.9470, 1.95, 1.95)$$

$$s = A^{-1}K = (0.7604, -0.0165, 0.0081, -0.0164, 0.0153, -0.1372, -0.1372, -0.1372) \quad (42)$$

using the equations below:

$$A = X'X, X = \begin{bmatrix} 1 & 0.85 & \cdots. & 0.22 \\ 1 & 0.62 & \cdots. & 0 \\ \cdots. & \cdots. & \cdots. & \cdots. \\ 1 & 0.55 & \cdots. & 0.22 \\ 1 & 0.20 & \cdots. & 0.56 \end{bmatrix} \quad (43)$$

$$Z = \left(\sum_{i=1}^{2592} z_i x_{i0}, \sum_{i=1}^{2592} z_i x_{i1}, \sum_{i=1}^{2592} z_i x_{i2}, \cdots, \sum_{i=1}^{2592} z_i x_{i7}\right)^T = (438.9808, 397.24, 339.81, 392.29, 354.40, 381.94, 382.35, 382.35) \quad (44)$$

$$K = \left(\sum_{i=1}^{2592} k_i x_{i0}, \sum_{i=1}^{2592} k x_{i1}, \sum_{i=1}^{2592} k_i x_{i2}, \cdots, \sum_{i=1}^{2592} k_i x_{i7}\right)^T = (1685.00, 932.8569, 1020.00, 1057.00, 1239.00, 419.2238, 419.2256, 419.2256) \quad (45)$$

The fuzzy logistic regression model is found as follows:

$$\tilde{W}_i = ln\left(\frac{\mu_i}{1-\mu_i}\right), i = 1,2,3,\cdots,2592 = (-3.9623, 0.7604) + (1.0909, -0.0165)x_{i1} + (1.0519, 0.0081)x_{i2} + (1.0964, -0.0164)x_{i3} + (0.9262, 0.0153)x_{i4} + (1.9470, -0.1372)x_{i5} + (1.95, -0.1372)x_{i6} + (1.95, -0.1372)x_{i7} \quad (46)$$

The possibilistic odds ratio of Observation 133 with a "high" score was calculated using the equation above as follows:

$$\tilde{W}_{133} = (-3.9623, 0.7604) + (1.0909, -0.0165) * 0.85 + (1.0519, 0.0081) * 0.77 + (1.0964, -0.0164)0.85 + (0.9262, 0.0153)0.85 + (1.9470, -0.1372) * 0.56 + (1.95, -0.1372) * 0 + (1.95, -0.1372) * 0.22 = (1.0134, 0.6447) \quad (47)$$

6. Discussion

The assessment criteria suggested by [29] the MSE and the MAE method were used to assess the proposed method [30]. $\tilde{Y}$ and $\hat{\tilde{Y}}$ are the real and estimated figures for the outputs, and Kim and Bishu's assessment criteria are as follows:

Table 4
Comparison of errors in estimation.

i. observation	Fuzzy output			Errors in estimation		
				Kim-Bishu	MSE	MAE
i. Low	−1.194	−0.497	0.199	1.5722	**0.7181**	0.8474
i. Very Low	−1.653	−0.925	−0.198	1.8682	1.3590	**1.1658**
i. High	−0.339	0.351	1.042	1.3129	**0.1358**	0.3685
i. Low	−1.39	−0.67	0.049	1.3226	1.028	**1.0142**
i. Critical	0.368	1.013	1.658	0.6242	**0.0069**	0.0834
i. Critical	0.889	1.490	2.090	1.7997	**0.2809**	0.5300
i. High	0.112	0.761	1.409	1.2701	**0.0026**	0.0512
i. Critical	0.369	1.014	1.659	1.8436	**0.0071**	0.0844

$$\phi = \int_{S_{\tilde{Y}} \cup S_{\hat{\tilde{Y}}}} \left| \tilde{Y}(y) - \hat{\tilde{Y}}(y) \right| dy \tag{48}$$

The $S_{\tilde{Y}}$ and $S_{\hat{\tilde{Y}}}$ parameters cover the bounds of $\tilde{Y}$ and $\hat{\tilde{Y}}$. The smaller ϕ value means better performance for the model.

Another criterion is that the real value was calculated using y_i while the estimated one was found using $\hat{\tilde{y}}_i$ (purified exact value) for the error in sample i. The smaller values mean that the estimation ability of the model is better.

$$MSE = \frac{1}{N} \sum \left(y_i - \hat{\tilde{y}}_i \right)^2 \tag{49}$$

MAE is that the real value was calculated using y_i while the estimated one was found using $\hat{\tilde{y}}_i$ (purified exact value) for the error in sample i. The smaller values mean that the estimation ability of the model is better.

$$MAE = \frac{1}{N} \sum \left| \hat{\tilde{y}}_i - y_i \right| \tag{50}$$

Furthermore, validity of the fuzzy regression model is demonstrated with different fitness functions. For fuzzy output data, the mean fuzzy credibility (MFC) is used.

The MFC is determined based on (51),

$$MFC = \sum_{i=1}^{N} \frac{\mu_{\tilde{y}}(y^c(i))}{\left| \tilde{y}^r(i) + \tilde{y}^l(i) \right|} \tag{51}$$

When MFC is high, the model generates high memberships with respect to all $y^c(i)$ and generates small uncertainty estimates. Despite the MFC, two other fuzzy criteria, namely average fuzzy spread (AFSC) and index of confidence (IC), are used to further validate the effectiveness and the performance of the fuzzy regressions, and they are as described as following:

AFSC given in (52) indicates the overall uncertainty predicted by the model. Unnecessary uncertainty is unlikely to be predicted by the model, when AFSC is small.

$$AFSC = \sum_{i=1}^{N} \left| \tilde{y}^r(i) - \tilde{y}^l(i) \right| \tag{52}$$

IC given in (53) evaluates the degree of variation of all collected samples to all fuzzy estimates with $i = 1, 2, \cdots, 2592$ which are the model outputs. The IC is similar to R-square in statistical regression, which indicates the capability of fitting the collected samples. When the IC is large, the fitting capability of the regression model is better [34].

$$IC = \frac{\sum_{i=1}^{N} \left(\tilde{y}\ c(i) - \tilde{y}^l(i) \right)^2 + \left(\tilde{y}\ c(i) - \tilde{y}^r(i) \right)^2}{\sum_{i=1}^{N} \left(y^c(i) - \tilde{y}^l(i) \right)^2 + \left(y^c(i) - \tilde{y}^r(i) \right)^2} \tag{53}$$

Kim and Bishu's method, MSE and MAE was compared for eight different observations in Table 4. Although the errors in Kim and Bishu's method were not significant, Kim and Bishu's method achieved a score of 1.7997 while MSE was 0.2809 for the "Critical" observation regarding the comparisons made for different linguistic thresholds. For the "Critical" observation, Kim and Bishu achieved 0.6242 while MSE gave 0.0069. Following the "Very Low" observation, Kim and Bishu achieved 1.8682 while MAE gave 1.1658. In other comparisons, MSE was substantially found to be more successful than the Kim and Bishu method and MAE. To sum up, the estimation ability of the model was found to be good as a result of the calculations made with MSE.

Table 5
Comparison of fitness functions.

Model for CVSS Dataset		
$\tilde{W}_i = (-3.9623, 0.7604) + (1.0909, -0.0165)x_{i1} + (1.0519, 0.0081)x_{i2} + (1.0964, -0.0164)x_{i3} + (0.9262, 0.0153)x_{i4} + (1.9470, -0.1372)x_{i5} + (1.95, -0.1372)x_{i6} + (1.95, -0.1372)x_{i7}$		
MFC	AFSC	IC
1.24	3.36	0.64

Table 6
The values of related CVSS database [33].

Observation No	Attack Vector (AV)	Attack Complexity (AC)	Privileges Required (PR)	User Interaction (UI)	Confidentiality (C)	Integrity (I)	Availability (A)	Severity Probability
1	0.85	0.77	0.27	0.62	0.56	0.22	0	Medium
2	0.85	0.77	0.85	0.85	0.56	0.22	0	Critical
3	0.85	0.77	0.68	0.85	0.56	0.22	0	High
4	0.85	0.77	0.5	0.85	0.56	0.22	0	High
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2498	0.85	0.77	0.85	0.62	0	0.56	0	Medium
2499	0.85	0.77	0.62	0.62	0	0.56	0	Medium
2590	0.85	0.77	0.27	0.62	0	0.56	0	Medium
2591	0.85	0.77	0.85	0.85	0	0.56	0	High
2592	0.85	0.77	0.68	0.85	0	0.56	0	High

Table 5 shows that for MFC the results are 1.24, MFC is expected to be large here, AFS is giving 3.36, AFSC is expected to take smaller value here, IC is giving 0.64, IC is expected to get large values. The proposed model is generally feasible and it can obtain better results when the three fuzzy criteria are taken into account.

7. Conclusions

Probability measurements and logarithmic conversion of probability rates regarding fuzzy outputs were performed to form an FLR model in this study. Evidence indicated that the classic logistic regression model may create uncertainty in the variables giving indefinite multiple categories. No assumptions regarding the explanatory variables are present in the logistic regression analysis. In cases with dual-response variables, Bernoulli probability distribution should be followed. However, no probability distribution can be considered for indefinite observations. The insufficiency of normal logistic regression in modeling indefinite dual observations, and the abundance of these observations in clinical and other studies, made this an interesting problem. The suggested model targeted the observations regarding the exact inputs and multiple fuzzy outputs. Compared to the studies on the FLR model, the suggested model was prepared based on the Least Squares Method formed by using Fuzzy-CVSS, which has certain advantages. For instance, the method models the relationships between the net inputs and fuzzy outputs (a linguistic term). Effective results were obtained by the proposed model when fitness functions are taken into account. Moreover, the estimated outputs are the fuzzy numbers representing the probability figures for the examined event. Ref. [31,32] modeled the fuzzy relationship between the net input–net output observations based on a probabilistic approach.

Regarding the logistic regression, fuzzy and indefinite observations were disregarded in establishing a model, while a relationship was defined between the fuzzy variables in the fuzzy logistic regression. Classic logistic regression cannot be implemented due to the fuzziness of CVSS severity in the suggested method. In addition, a new field was used for the first time, which is different to the logistic regression model in a fuzzy environment. However, the suggested model can be expanded in cases where both explanatory variables and response variables are fuzzy. Moreover, different parameter estimation or model assessment methods can be used to improve the model. In conclusion, with the use of linguistic terms to assess the common vulnerabilities, the expectation is that it will assist the administrators of organizations on a broader scale with better error tolerance levels in taking action against the most critical vulnerabilities.

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.

References

[1] Mell P, Scarfone K, Romanosky S. A Complete Guide to the Common Vulnerability Scoring System Version 2.0. Forum of Incident Response and Security Teams (FIRST); 2007.
[2] CVSS. Common Vulnerabilities Scoring System (CVSS), https://www.first.org/cvss/v3.0/specification-document [27.12.2019].
[3] Khazaei A, Ghasemzadeh M, Derhami V. An automatic method for CVSS score prediction using vulnerabilities description. J Intell Fuzzy Syst 2015;30:89–96.
[4] Tanaka H. Linear regression analysis with fuzzy model. IEEE Trans Syst, Man, Cybernetics 1982;12(6):903–7.
[5] Tanaka H, Uejima S, Asai K. Linear regression analysis with fuzzy model. IEEE Trans Syst, Man, Cybernetics 1982;12:903–7.
[6] Hojati M, Bector CR, Smimou K. A simple method for computation of fuzzy linear regression. Eur J Oper Res 2005;166(1):172–84.
[7] Modarres M, Nasrabadi E, Nasrabadi MM. Fuzzy linear regression models with least square errors. Appl Math Comput 2005;163(2):977–89.
[8] Diamond P. Fuzzy least squares. Inf Sci 1988;46(3):141–57.
[9] Celmiņš A. Least squares model fitting to fuzzy vector data. Fuzzy Sets Syst 1987;22(3):245–69.
[10] Ming M, Friedman M, Kandel A. General fuzzy least squares. Fuzzy Sets Syst 1997;88(1):107–18.
[11] Yang M-S, Lin T-S. Fuzzy least-squares linear regression analysis for fuzzy input–output data. Fuzzy Sets Syst 2002;126(3):389–99.
[12] Dondo MG. A Vulnerability Prioritization System Using A Fuzzy Risk Analysis Approach. In: Proceedings of The Ifip Tc 11 23rd International Information Security Conference, Boston, MA; 2008, p. 525–40.
[13] Anikin V. Using fuzzy logic for vulnerability assessment in telecommunication network. In: 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM); 2017, p. 1–4.
[14] Houmb SH, Franqueira VNL. Estimating ToE Risk Level Using CVSS. In: 2009 International Conference on Availability, Reliability and Security; 2009, p. 718–25.
[15] Pourahmad S, Ayatollahi SMT, Taheri SM, Agahi ZH. Fuzzy logistic regression based on the least squares approach with application in clinical studies. Comput Math Appl 2011;62(9):3353–65.
[16] Sohn SY, Kim DH, Yoon JH. Technology credit scoring model with fuzzy logistic regression. Appl Soft Comput 2016;43:150–8.
[17] Tanaka H. Fuzzy data analysis by possibilistic linear models. Fuzzy Sets Syst 1987;24(3):363–75.
[18] Tanaka H, Watada J. Possibilistic linear systems and their application to the linear regression model. Fuzzy Sets Syst 1988;27(3):275–89.
[19] Shakouri GH, Nadimi R. A novel fuzzy linear regression model based on a non-equality possibility index and optimum uncertainty. Appl Soft Comput 2009;9(2):590–8.
[20] Kao C, Chyu C-L. A fuzzy linear regression model with better explanatory power. Fuzzy Sets Syst 2002;126(3):401–9.

[21] Sakawa M, Yano H. Fuzzy linear regression analysis for fuzzy input-output data. Information Sci 1992;63(3):191–206.
[22] Chan KY, Engelke U. Varying spread fuzzy regression for affective quality estimation. IEEE Trans. Fuzzy Syst. 2017;25(3):594–613. doi: https://doi.org/10.1109/TFUZZ.2016.2566812.
[23] Liu Y, Chen Y, Zhou J, Zhong S. Fuzzy linear regression models for QFD using optimized h values. Eng Appl Artif Intell 2015;39:45–54.
[24] Jiang HM, Kwong CK, Ip WH, Chen ZQ. Chaos-based fuzzy regression approach to modelling customer satisfaction for product design. IEEE Trans Fuzzy Syst 2013;21(5):926–36.
[25] Muzzily S, Ruggieri A, Baets BD. A comparison of fuzzy methods for the estimation of the implied volatility smile function. Fuzzy Sets Syst 2015;226:131–43.
[26] Chan Y, Engelke U. Fuzzy regression for perceptual image quality assessment. En. Appl Artif Intell 2015;43:102–10.
[27] Wang Y-M, Yang J-B, Xu D-L, Chin K-S. On the centroids of fuzzy numbers. Fuzzy Sets Syst 2006;157(7):919–26.
[28] Namdari, Yoon JH, Abadi A, Taheri SM, Choi SH. Fuzzy logistic regression with least absolute deviations estimators. Soft Computing 2015;19(4):909–17.
[29] Kim B, Bishu RR. Evaluation of fuzzy linear regression models by comparing membership functions. Fuzzy Sets Syst 1998;100(1):343–52.
[30] Gao Y, Lu Q. A fuzzy logistic regression model based on the least squares estimation. Computational Appl Mathematics 2018;37(3):3562–79.
[31] Nagar P, Srivastava S. Adaptive fuzzy regression model for the prediction of dichotomous response variables using cancer data: a case study. J Appl Math, Statistics Informatics 2008;4:183–91.
[32] Dom RM, Zain R, Abdul Kareem S, Abidin B. An adaptive fuzzy regression model for the prediction of dichotomous response variables. In: 15th Conference on Computational Science and Applications. Malaysia; 2007, p. 14–9.
[33] FIRST, CVSS/Calculator, https://www.first.org/cvss/calculator/3.0 [27.12.2019].
[34] Chan KY, Lam H-K, Yiu CKF, Dillon TS. A flexible fuzzy regression method for addressing nonlinear uncertainty on aesthetic quality assessments. IEEE Trans Syst Man Cybern Syst 2017;47(8):2363–77.