RAIRO-Oper. Res. 56 (2022) 4083–4111
https://doi.org/10.1051/ro/2022189

RAIRO Operations Research
www.rairo-ro.org

AN APPLICATION OF FUZZY LOGISTIC REGRESSION FOR PREDICTING CVSS SEVERITY CATEGORY OF INDUSTRIAL CONTROL SYSTEMS

Ahmet Murat Dere* and Mehmet Kabak

Abstract. Cybersecurity is rapidly gaining significance due to growing use of computers in daily life and business sectors. Likewise, industrial sector has also become more vulnerable to cyber threats exclusively with the onset of Industry 4.0, which is a digital transformation evolved with industrial control systems (ICS). Nowadays industrial organizations aim to build capacity towards protection of ICS to be cybersafe. To assess the effects of vulnerabilities in ICS, organizations utilize Common Vulnerability Scoring System (CVSS), which calculates severity categories/scores. In this study, we implemented a prediction model for CVSS vulnerability categorization of ICS. Although there exist many applicable methods to use in data analysis paradigm such as statistical regression, cluster and classification analysis, the categorical form of CVSS data based on verbal statements and the failure to satisfy basic statistical assumptions for classical models motivated us to focus on implementation of fuzzy logistic regression (FLR) model, which is one possible alternative method. We chose the FLR method to explore that it is applicable to ICS vulnerability data. Furthermore, the model was improved by employing metaheuristic algorithms to optimize the spread of fuzzy numbers representing input variables. This study is expected to contribute to practical application of vulnerability categorization of ICS.

Mathematics Subject Classification. 62A86.

Received February 23, 2022. Accepted October 22, 2022.

1. Introduction

Digital Era has profoundly affected our daily life. Now, we live in a digitally interconnected world and most of these conveniences have already irreversibly been integrated into our life. Moreover, traditional methods of education, production, health and commerce have undergone extensive changes to take advantage of the innovations brought by the Digital Era. This deep impact of digitalization also triggered Industry 4.0 in the industrial sector. Industry 4.0 is defined as the fourth generation of the industrial revolution and digital industrial transformation in parallel with developments in informatics, ICS, big data, internet of things (IoT), artificial intelligence (AI), cloud computing, machine learning. By integrating these digital technologies into various processes and making better decisions by analyzing data from the supply chain, Enterprise Resource Planning (ERP), customer service and other systems, Industry 4.0 is improving overall efficiency in operations management [1]. However, all these systems inevitably face an inherent threat created by digital technology itself, and that is the cyber threat. A

Keywords. ICS, CVSS, Fuzzy Logistic Regression, metaheuristic algorithms.

Industrial Engineering Department, Gazi University, Ankara, Turkey.
*Corresponding author: ahmet.murat.dere@gazi.edu.tr

cyber threat source may exploit a vulnerability, a weakness in an information system, system security procedures, internal controls, or implementation [2]. A flaw or weakness could be exercised (accidentally triggered or intentionally exploited) and result in a security breach or a violation of the system's security policy [3]. A weakness in the computational logic (*e.g.*, code) found in software and hardware components, when exploited, results in a negative impact to confidentiality, integrity, or availability (of the system), which may likely cause monetary losses, disruption of processes, digital theft, and damage to institutional reputation. Mitigation of vulnerabilities typically involves coding changes, but could also include specification changes or even specification deprecations (*e.g.*, removal of affected protocols or functionality in their entirety) [4]. In this context, all information systems including countries' critical infrastructure, ICS, personal smartphones, and even home-use robot vacuum cleaners are vulnerably exposed to the ever-increasing risk of these threats. Since cybersecurity preparedness entails a close and up-to-date support from various providers, organizations – large or small – include cybersecurity costs in institutional budget, which has made cybersecurity a huge market. In 2004, the global cybersecurity market was worth just $3.5 billion, but with cybercrime predicted to cost the world $10.5 trillion annually by 2025 (up from $3 trillion a decade ago and $6 trillion in 2021), the cumulative global spending on cybersecurity products and services for the five-year period from 2021 to 2025 is predicted to have a market value of $1.75 trillion [5]. Therefore, organizations, to sustain a desired level of operational effectiveness, rigorously seek ways to promote security posture by identifying potential vulnerabilities in their systems to prioritize cybersecurity investment. To achieve this goal, organizations predominantly use the Common Vulnerability Scoring System (CVSS) to assess the severity of potential vulnerabilities (*i.e.*, None, Low, Medium, High and Critical). CVSS provides a standard metric and captures the principal technical characteristics of software, hardware and firmware vulnerabilities. CVSS was first designed by The National Infrastructure Advisory Council (NIAC), a US government advisory council, in 2003, then the Council chose the Forum of Incident Response and Security Teams (FIRST) to develop and manage CVSS. Vulnerability data feeds for analysis are available at the National Vulnerability Database [6] kept up to date by the National Institute of Standards and Technology (NIST), which is a part of the U.S. Department of Commerce. CVSS calculates a base severity category or a score of a vulnerability, based on some categorical input variables values defined by expert evaluation/human assessment. Typically, the vendor who announces the vulnerability provides certain details to create CVSS scores. If the vendor declines to provide certain details, National Vulnerability Database (NVD) analysts assign CVSS scores using a worst-case approach. If a vendor provides no details about a vulnerability, NVD will score that vulnerability as a 10.0 (the highest rating) [7].

The vulnerability categories/scores of ICS are also determined by using CVSS. Furthermore, the Cybersecurity and Infrastructure Security Agency (CISA), a part of the U.S. Department of Homeland Security, promotes a cohesive effort between government and industry to improve the cyber security posture of ICS by providing an online database for advisories and alerts about current cyber security issues, vulnerabilities and exploits. CISA also uses CVSS to identify the category of security vulnerabilities of ICS [8].

ICS vulnerability data based on CVSS is in categorical form. The values assigned by an expert and both input and output variables are expressed in linguistic terms.

There are many methods to analyze categorical data. In this study, on the other hand, we emphasize on fuzzy regression analysis which is an extension of (or an alternative for) the classical regression analysis in which some elements of the model are represented by fuzzy numbers [9]. The fuzzy models based on fuzzy set theory provides a proper substitution for modeling to handle the uncertainty and irrepressible imprecision in the data. Fuzzy regression approach is also a promising alternative, when distributional assumptions of the underlying regression model are not satisfied or they cannot be tested. There are many situations in practical applications where the data cannot be measured as crisp quantities, since information in data is often irrepressibly imprecise, vague, incomplete, verbal/linguistic, and qualitative similar to the case of CVSS vulnerability data. As can be seen in the literature review section, fuzzy modeling approaches provide a promising alternative techniques for dealing with such situations and fuzzy regression models have been extensively used in categorical data analysis [10]. All of those considerations motivated us to conclude that Fuzzy Logistic Regression (FLR) is the most viable alternative method for analyzing the ICS vulnerability data based on CVSS.

There has been much research on CVSS, Fuzzy Regression (FR) and FLR, however, this study is the first attempt to use FLR method to demonstrate that fuzzy input and fuzzy output FLR models can be applicable to ICS vulnerability data based on CVSS. The applied model is also improved by applying metaheuristic algorithms for optimizing the spread of fuzzy numbers representing input variables. Section 2 reviews relevant literature. Section 3 introduces data, briefly explains the basic concept of fuzzy set theory and the applied methodology, and conducts FLR models. Section 4 discusses the performance of the model and the results. Section 5 concludes the study.

2. Literature review

In this section, we first review some approaches for CVSS prediction models. We then focus on the previous studies in the FR and FLR.

2.1. General CVSS prediction models

CVSS has specific equations/formulas defined to calculate the score based on CVSS metrics [11]. However, there are numerous studies in the literature that explore and propose various alternative models. Dondo [12] presented a fuzzy inference system that ranks vulnerabilities by using fuzzy rules. They argued that the approach has more meaningful vulnerability prioritization values than the severity level calculated by the popular CVSS. Their approach was shown to be capable of ranking vulnerabilities over networks and organizations, which CVSS scores cannot provide comparable rankings for such cases. Anikin [13] suggested a risk assessment model using the fuzzy inference method for these fuzzy rules to describe experts' knowledge. They concluded that their method has some advantages over CVSS. Gencer and Başçiftçi [14] presented a novel model that explored how CVSS vulnerability could be calculated using linguistic terms by an FLR model only with crisp (non-fuzzy) input and fuzzy output. They did implement an FLR model, rather than classic logistic regression model, due to the fuzzy nature of CVSS data. The performance was evaluated by a several different criteria and they obtained effective results.

2.2. FR models

Regression analysis is a powerful statistical method for estimating the relationships between a dependent variable and one or more independent variables. It is extensively used in many areas, such as engineering, finance and economics, biology, and environmental sciences. Statistical regression models can actually cope with randomness in the data, and only use observations which are in numeric form with exact/crisp values. On the other hand, it is very common in real life that the data is defined/collected as linguistic/verbal terms (such as low, medium, high) based on subjective human assessment. This naturally yields uncertainty or vagueness caused by imprecise boundaries between the different levels of categories assigned to linguistic terms [15]. In this case, rather than randomness, uncertainty should be taken into consideration. Besides, the basic assumptions (such as normality, the identical distribution of the error terms) for statistical models do not hold [16]. In such situations, fuzzy models are an alternative method which may perform better in capturing and extracting useful information from imprecise data [17].

FR was introduced by Tanaka *et al.* [18]. Classical regression is based on probability theory, whereas fuzzy regression is based on Zadeh's fuzzy set theory [19]. There are two main approaches for FR. The first one is the probabilistic approach by Tanaka *et al.* [18]. They developed a linear programming model to minimize the total spread of each fuzzy coefficient and the total vagueness of the estimated output. The second one is the fuzzy least squares approach by Diamond [20] that is based on minimizing the distance between fuzzy observed and estimated outputs. Unfortunately, Tanaka's approach only allows the responses to be fuzzy numbers and the predictors to be crisp [21]. In the literature, they are many studies that applied various methods of FR successfully. Some studies considered input variables to be crisp and output are fuzzy, while other studies analyzed FR with both fuzzy input and outputs.

Yoon and Choi [22] were able to provide a new least-squares approach for fuzzy regression through one compact formula derived from triangular fuzzy matrices and they defined also fuzzy matrix new operations. This approach opened a new perspective for methods in fuzzy regression since it became possible to express estimators in one formula that makes it easy to prove optimal or asymptotic properties. In their study, they considered a fuzzy linear regression model with fuzzy input and fuzzy output data.

Yoon and Grzegorzewski [23], as a follow-on study, explained details and properties of the least-squares approach based on Yoon and Choi [22]. It was also shown that a fuzzy least-squares approach is originally a fuzzy generalization/extension of ordinary (crisp) least-squares. This method was also successfully applied by Sohn *et al.* [21] in an FLR model with fuzzy input and fuzzy binary output.

Chukhrova and Johannssen [10] focused on presenting a comprehensive systematic review on the topic of FR analysis.

2.3. FLR models

FLR, an extension of FR, models the relationship between fuzzy independent variables and fuzzy categorical (multinomial/ordinal or binary) dependent variables. Although it is relatively new, there exists a considerable amount of literature available on FLR models.

Pourahmad *et al.* [16] introduced a new term called "possibilistic odds", which is the possibility rate of success to non-success. Due to the imprecise nature of the output variable, no underlying probability distribution can be assumed. Hence, instead of using classical logistic regression, they introduced an FLR model based on possibilistic odds. In their model, the input was crisp, and the output was fuzzy binary. They also proposed a new goodness-of-fit criteria called Mean Degree of Memberships (MDM). This approach was used in many other studies.

Pourahmad *et al.* [9] proposed an FLR approach based on the least-squares method. They considered a crisp input and fuzzy multinomial output FLR model. To evaluate the proposed model, they adopted Mean Capability Index (MCI) presented by Taheri and Kelkinnama [22].

Sohn *et al.* [21] used the fuzzy least-square estimation method to fit an FLR model for technology credit scoring, which consists of fuzzy input and fuzzy binary output variables. They successfully applied the least-squares approach proposed by Yoon and Choi [22]. They used sensitivity, specificity, accuracy, and a ROC Curve to evaluate the performance of the proposed model. They also compared the fitted FLR model with an ordinary/classical logistic regression.

Yapıcı Pehlivan and Yonar [19] introduced an FLR model with crisp input and fuzzy binary output variables based on possibilistic odds presented by Pourahmad *et al.* [16]. They used the least-squares approach of Diamond [20] to estimate the parameters of the FLR model. As for the goodness-of-fit criteria, they also used sensitivity, specificity and accuracy to compare their FLR model with an ordinary/classical logistic regression.

Atalik and Senturk [25] introduced an FLR model based on Tanaka's regression model in which the objective function is improved. An application is performed on a crisp input and fuzzy binary output data set. They evaluated their model using the MDM criteria.

Namdari *et al.* [26] constructed a crisp input and fuzzy multinomial output FLR model using least-squares approach to examine the effect of folic acid on appetite in children. The results of the FLR and a statistical ordinal logistic regression (OLR) was compared.

Namdari *et al.* [27] proposed a new estimator for FLR with crisp input and fuzzy ordinal output. This estimator was called "Least Absolute Deviations (LAD)", and the results were compared with the typical least-squares estimation (LSE) method. They presented two new goodness-of-fit indices called Measure of Performance based on Fuzzy Distance (M_p) and Index of Sensitivity for outliers (I_s).

Gencer and Başçiftçi [14] applied an FLR model based on least-squares approach. The input variables were crisp, and output was fuzzy multinomial. As for assessment criteria, they adopted the Kim and Bishu [28]'s criterion, mean squared error (MSE) and mean absolute error (MAE). In their model CVSS data is used. For further study, they suggested a fuzzy input-fuzzy output model.

Salmani *et al.* [29] proposed a forward variable selection method for FLR for fuzzy/crisp input and fuzzy multinomial output. They used the least-squares estimation method introduced in Xu and Li [30] as an extension of traditional least-squares approach in a fuzzy environment as in Diamond [20]. They measured the goodness-of-fit of the stepwise modeling with fuzzy extensions of MSE, Akaike Information Criteria (AIC), Mallows Measure (C_p).

Salmani *et al.* [31] applied a least-squares method to construct an FLR with crisp input and fuzzy multinomial output. The least-squares approach was the same as in Salmani *et al.* [29] which was proposed by Xu and Li [30] as an extension of traditional least-squares approach in a fuzzy environment as in Diamond [20]. MCI was used to measure the goodness-of-fit.

Mustafa *et al.* [17] developed a least-squares FLR model with crisp input and fuzzy binary output. The model was evaluated by MCI. This study considered a trapezoidal membership function which is different from most studies in FLR literature.

Gao and Lu [15] developed a least-squares FLR model with crisp/fuzzy input and fuzzy multinomial output. They proposed a fuzzy adjustment error term. For evaluation criteria, Kim–Bishu and MCI were adopted.

Nikbakht and Bahrampour [32] used an FLR model with crisp input and fuzzy binary output to determine predictive survival factors of breast cancer patients. The performance was determined in terms of MDM.

Behnampour *et al.* [33] considered an FLR model with crisp input and fuzzy multinomial output to predict the severity of autism. The fuzzy model parameters were estimated by LSE and LAD with an approach presented in Kelkinnama and Taheri [34]. The results of LSE and LAD M_p was compared in terms of M_p criterion.

Bennaser [35] focused on developing an FLR model with crisp input and fuzzy multinomial output for DNA methylation data. He compared three methods, weighted average logistic regression (WALR), OLR, and FLR based on the correct classification rates.

Anggraeni *et al.* [36] applied FLR to predict Dengue fever outbreak with fuzzy input variables a fuzzy multinomial output. For estimation of parameters, maximum likelihood estimator (MLE) was used. They compared the performance of the FLR model with Neural Network, Random Forest, and Naive Bayes approaches.

This study and the main relevant previous studies are compared in Table 1 in terms of similar approaches in methodology available in the relevant literature.

According to Table 1, a closer look at the relevant literature reveals a gap: an FLR model with fuzzy input and fuzzy multinomial/ordinal output. Although, crisp input and fuzzy output models have drawn much attention, FLR models with fuzzy input and fuzzy multinomial/ordinal output are quite limited. Especially, for CVSS vulnerability data, only one study [14] exists in the relevant literature, with a crisp input and fuzzy multinomial output. The ICS vulnerability data, based on CVSS, the input (independent) and output (dependent) variables are in ordinal/multinomial categorical variable form and the values assigned are defined by linguistic terms (*i.e.*, low, medium, high, etc.), which intrinsically holds uncertainty caused by subjective human assessment. Within this regard, an FLR model to analyze ICS vulnerability data is an appropriate method. Hence, an FLR model based on least-squares with fuzzy input and fuzzy multinomial output model applied on ICS vulnerability data is assumed to be a contribution in terms of application and addresses the gap in the relevant literature.

This study demonstrates that FLR models can be applied to fuzzy input and fuzzy multinomial/ordinal output ICS vulnerability data based on CVSS. For parameter estimation, the fuzzy least-squares estimate concept proposed in Yoon and Choi [22], Yoon and Grzegorzewski [23] is used. The form of data, in our case fuzzy input and fuzzy multinomial/ordinal output, and the least-squares method for parameter estimation are different from the approaches proposed in Gencer and Başçiftçi [14] and Sohn *et al.* [21], which are the closest studies to our methodology. Moreover, after estimating the parameters of FLR using initial spreads of fuzzy input variables, the model is improved by applying metaheuristic algorithms to optimize the values of the spreads. To the best of our knowledge, no prior studies have examined these issues. Therefore, this is the first study in the literature to investigate the applicability of the FLR model to ICS data and optimize the spread of fuzzy numbers using metaheuristic algorithms.

TABLE 1. Current study *vs.* the previous studies in the relevant literature.

Authors	Method		Form of data		Performance criteria
	Possibilistic	Least-Squares	Input	Output	
Pourahmad *et al.* [16]	Tanaka *et al.* [18]		Crisp	Fuzzy binary	MDM MSE
Pourahmad *et al.* [9]		Diamond [20]	Crisp	Fuzzy multinomial	MCI
Yapıcı Pehlivan and Yonar [19]		Diamond [20]	Crisp	Fuzzy Binary	Sensitivity, Specificity, Accuracy.
Atalik and Senturk [25]	Tanaka *et al.* [18]		Crisp	Fuzzy Binary	MDM
Namdari *et al.* [26]		Diamond [20]	Crisp	Fuzzy ordinal	Comparison with ordinary logistic regression
Mustafa *et al.* [17]		Diamond [20]	Crisp	Fuzzy Binary	MCI
Salmani *et al.* [31]		Xu and Li [30]	Crisp	Fuzzy Multinomial	MCI
Gao and Lu [15]		Diamond [20]	Crisp/ Fuzzy	Fuzzy Multinomial	Kim–Bishu MCI
Nikbakht and Bahrampour [32]	Tanaka *et al.* [18]		Crisp	Fuzzy Binary	MDM
Bennaser [35]		Diamond [20]	Crisp	Fuzzy Multinomial	True Positive and False Positive Rates
Gencer and Başçiftçi [14]		Diamond [20]	Crisp	Fuzzy Multinomial	MSE, MAE, Kim–Bishu
Sohn *et al.* [21]		Yoon and Choi [22]	Fuzzy	Fuzzy Binary	Sensitivity, Specificity, Accuracy, ROC Curve
This study		Yoon and Choi [22]	Fuzzy	Fuzzy Multinomial	Kim–Bishu MCI MDM Accuracy

3. METHODOLOGY

This section presents the ICS vulnerability data based on CVSS, briefly reviews the underlying concepts and basic theory and introduces the FLR model application.

3.1. Data

ICS Data, which uses the CVSS metrics, is a special subset of CVSS vulnerability data in terms of ICS and is available on the CISA online databases [8]. Therefore, explaining the data structure of CVSS will also cover the definition of ICS vulnerability data. However, when it comes to comparison of ICS and CVSS scores, we showed with a t-test that there is a statistically significant difference between ICS and CVSS data, and also the mean score of ICS is larger than that of CVSS. Technically, it means that ICS is more vulnerable to cyber threats, and it forms a different group than general CVSS, which is for all computer systems. CVSS is composed of three metrics groups: Basic, Temporal and Environmental. The "Base" metric group represents the intrinsic characteristics of a vulnerability that are "constant" over time and across user environments. The "Temporal" metric group reflects the characteristics of a vulnerability that may change over time but not across user environments. For example, the presence of an exploit code could increase the CVSS score today, after some

TABLE 2. A sample portion of the dataset.

Obs number	Base category	AV	AC	PR	UI	Scope	C	I	A	Base score
1	HIGH	LOCAL	LOW	NONE	REQUIRED	UNCHANGED	HIGH	HIGH	HIGH	7.8
2	HIGH	NETWORK	LOW	NONE	NONE	UNCHANGED	NONE	NONE	HIGH	7.5
3	CRITICAL	NETWORK	HIGH	NONE	NONE	CHANGED	HIGH	HIGH	HIGH	9
4	HIGH	LOCAL	LOW	LOW	NONE	CHANGED	HIGH	HIGH	HIGH	8.8
5	HIGH	NETWORK	LOW	NONE	NONE	UNCHANGED	HIGH	NONE	NONE	7.5
6	HIGH	NETWORK	LOW	LOW	NONE	UNCHANGED	HIGH	HIGH	HIGH	8.8
7	MEDIUM	NETWORK	HIGH	NONE	REQUIRED	CHANGED	NONE	HIGH	NONE	6.1
8	HIGH	LOCAL	HIGH	LOW	NONE	CHANGED	NONE	HIGH	HIGH	7.5
9	CRITICAL	NETWORK	HIGH	NONE	NONE	CHANGED	HIGH	HIGH	HIGH	9
10	MEDIUM	NETWORK	LOW	HIGH	REQUIRED	UNCHANGED	HIGH	HIGH	HIGH	6.8

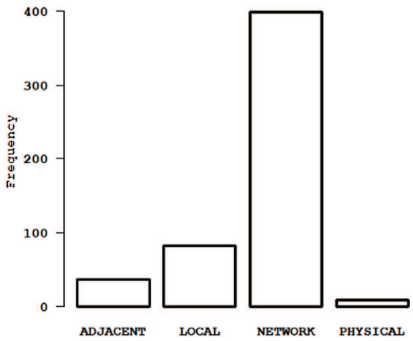

FIGURE 1. Attack Vector.

time when the creation of an official fix or patch would decrease the score. The "Environmental" metric group represents the characteristics of a vulnerability relevant and unique to a particular user's specific environment. Generally, only the score calculated on the Base metric is used for analysis. The "Base" metric consists of five basic components (total of eight categorical variables) to calculate a score: Attack Vector (AV), Attack Complexity (AC), Privileges Required (PR), User Interaction (UI), Scope, Confidentiality (C), Integrity (I), and Availability (A) (Tab. 2).

In this study, a set of ICS vulnerability based on CVSS (version 3.0/3.1) data containing 525 samples (Figs. 1–11) was downloaded from the CISA database. This dataset was split into train set ($n = 365$, 70% of all data) and test set ($n = 160$, 30% of all data). Another set of data containing 100 samples was separately collected from the same source as the out-data set to unbiasedly evaluate the performance of the final models with out-of-sample data.

Attack Vector (Network, Adjacent, Local, and Physical) reflects the context by which vulnerability exploitation is possible. Network means remote exploit over the internet. Physical means that the attack requires the attacker to physically touch or manipulate the vulnerable component.

Attack Complexity (High, Low) describes the conditions beyond the attacker's control that must exist to exploit the vulnerability. It is "High" when it requires the attacker to invest in some measurable amount of effort in preparation or execution against the vulnerable component before a successful attack can be expected. The lower the "Attack Complexity" is, the greater "the score" is, since the attack is easy to execute.

Privileges Required (None, Low, High) describes the level of privileges (*i.e.*, admin or user) an attacker must possess before successfully exploiting the vulnerability. If no privileges are required, the score is higher since anyone without a privilege could exploit the system.

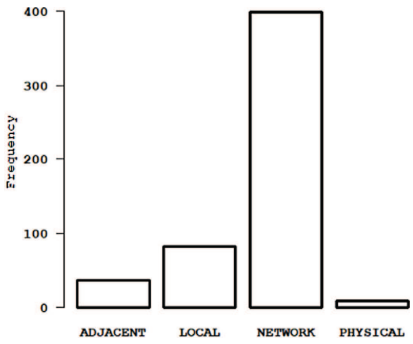

FIGURE 2. Attack Complexity.

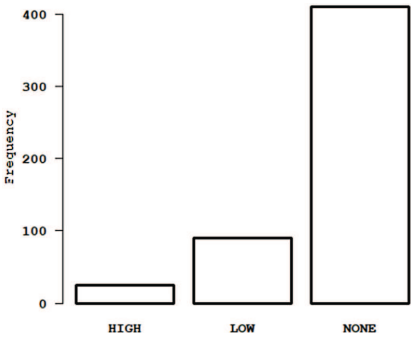

FIGURE 3. Privileges Required.

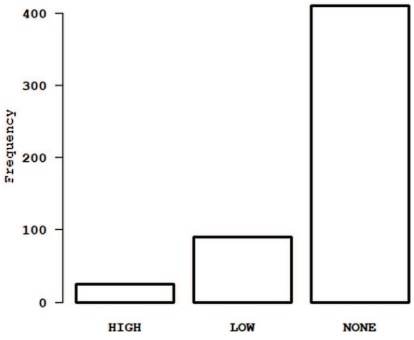

FIGURE 4. User Interaction.

User Interaction (None, Required) captures the requirement for a user, other than the attacker, to interact with the system such as installing an application. If no "User Interaction" is required, the score becomes larger.

Scope (Changed, Unchanged) captures whether a vulnerability in one vulnerable component impacts resources in components beyond its security scope. Typically, in a certain organization, all devices or components (*e.g.*, files, CPU, memory) are under a single security authority meaning under the same single jurisdiction of a security Scope. If an attack affects resources beyond the Scope of that organization (*e.g.*, cloud databases remotely accessed) and under control of another Scope, it means that Scope is changed. The "Scope" parameter

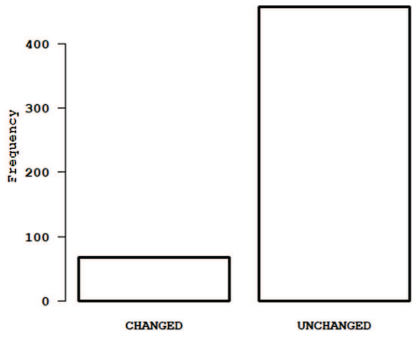

FIGURE 5. Scope.

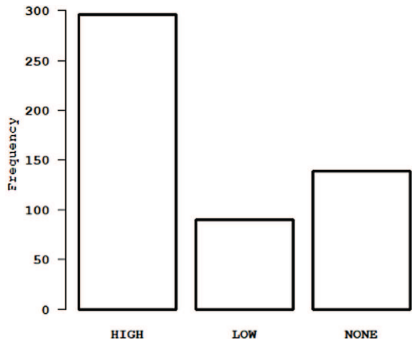

FIGURE 6. Confidentiality.

was not used in the model not to create redundancy or correlation due to its definition in the CVSS Specification Document, which states that different levels of Scope directly change the numeric value of Privileges Required variable, so the effect of Scope is already covered by "Privileges Required".

Confidentiality (High, Low, None) measures the impact on the confidentiality of the information resources managed by a software component due to a successfully exploited vulnerability. It is "High" when there is a total loss of confidentiality (*e.g.*, an attacker steals the administrator's password).

Integrity (High, Low, None) measures the impact on the integrity of a successfully exploited vulnerability. Integrity attains the value "High", when the attacker can modify any files protected by the impacted component. In this case, the information stored on the system is no longer trustworthy.

Availability (High, Low, None) measures the impact on the availability of the impacted component resulting from a successfully exploited vulnerability. It is "High" when the attacker has the ability to deny some availability (*e.g.*, the attacker cannot disrupt existing connections, but can prevent the new connections, or after repeated exploitation the attacker causes a service to become completely unavailable due to all memory is used) [11,14].

There is also an online calculator on FIRST's website where it can be readily seen that an expert can assign values to metrics by selecting categories of input variables and the output is calculated using specific equations. These input data are basically in binary/ordinal/multinomial form. When the values for Base metrics are assigned by an analyst (human assessment), the Base equation defined in [11] computes a base score ranging from 0.0 to 10.0, and this rating is then transformed into qualitative categories (None, Low, Medium, High, and Critical). The cut-off points for borderlines between categories are also specified by human judgment, so the output variable is imprecise and there exists uncertainty (and fuzziness) its nature. Therefore, no statistical

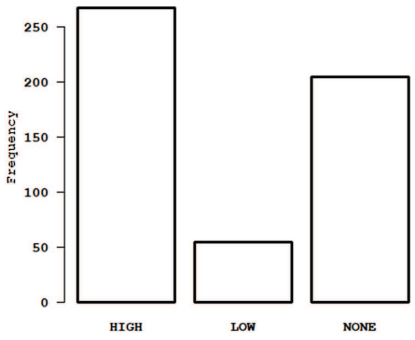

FIGURE 7. Integrity.

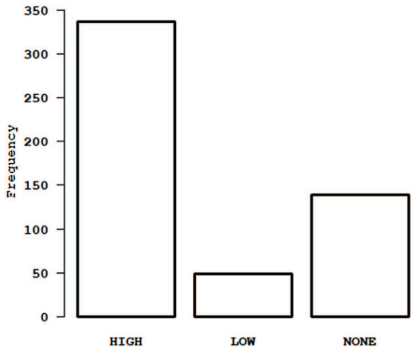

FIGURE 8. Availability.

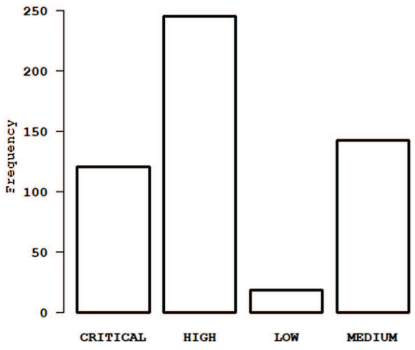

FIGURE 9. Base Category.

distribution can be fitted to the output data due to innate uncertainty which is not caused by randomness. Besides, when experts are expressing their assessment in linguistic/verbal terms, their opinions may vary, different values could be allotted while inspecting the very same vulnerability case. So, the classical statistical models can't address this issue. Uncertainty may be represented in exact/crisp value, but this simplification may cause information loss in prediction models. In this case, fuzzy regression model based on fuzzy set theory can account for the irreducible uncertainty inherent in data induced by human judgment. Indeed, the categorical form in data requires a statistical logistic regression model for data analysis while the linguistic nature of the data

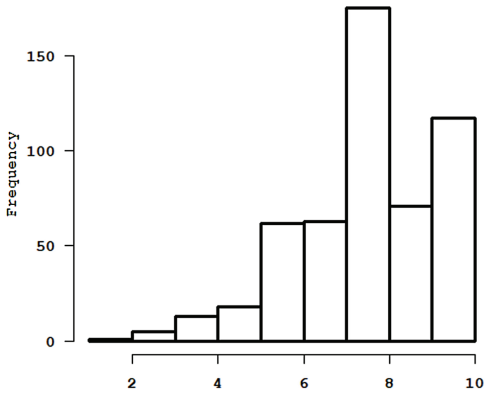

FIGURE 10. The histogram of Base score (continuous).

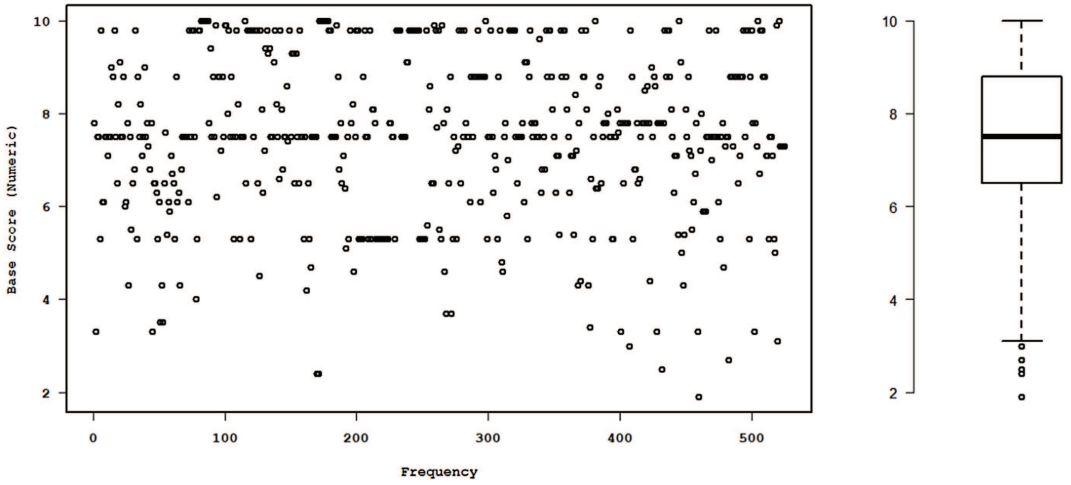

FIGURE 11. The scatter plot of Base score (along with its boxplot on the right).

entails methods on fuzzy set theory. Therefore, in the current study, an FLR model, which is a generalization of logistic regression in a fuzzy environment using least-squares approach, is fitted to the ICS data.

For modeling since input and output variables are linguistic terms, they are characterized by triangular fuzzy numbers (TFN). Due to the computational simplicity in arithmetic operations, TFN is the most widely used fuzzy number type in literature and has a common range of uses [19].

3.2. Basic introduction to fuzzy set theory and FLR

This section briefly reviews the basic definitions, operations on fuzzy sets, the underlying concepts of the FLR model and metaheuristic algorithms.

Fuzzy logic (FL), first introduced by Zadeh [37], can be defined as a mathematical model to study and define uncertainties, which is very prevalent in real life. FL systemically deals with imprecision and uncertainty, efficiently evaluates linguistic/verbal categorical terms defined by subjective human assessment, which are represented by fuzzy numbers, and successfully extracts useful information. Prediction models based on classical models are not feasible, since the basic statistical assumptions are not met for the uncertain/imprecise data.

Much research on FL as a forecast model has been conducted, especially in engineering. FL is a process of mapping an input space onto an output space using membership functions (MF) and linguistically specified rules [38]. The fundamentals of a set in FL are basic and simple: "concurrent partial membership or belonging to different subsets of the universal set instead of full belonging to a single set" and "a gradual membership (or being element) of a set". The gradual membership degree on the interval $[0,1]$ is numerically associated with an element value defined by MFs. An MF, the selection of which also generally depends on human assessment, can be in the form of different shapes for different kinds of fuzzy sets, such as triangle, and trapezoid, Gaussian, bell, and sigmoid. By FL, the degree of trueness of a statement can simultaneously be a member of the "totally true set" and "totally false set" with different degrees of membership respectively, which is not limited to just the classical Boolean logic values (true or false) [39]. The ideas of Aristotle's classical two-valued Boolean logic and Łukasiewicz's three-valued logic are comprehensively covered by the concept of FL. The FL enabled mathematical prediction models based on linguistic/verbal/qualitative statements (such as low, medium, high) in natural language, which triggered a new approach in approximate human reasoning. The basic definitions of FL are as follows:

Definition 3.1. A fuzzy set $\widetilde{A}$ defined in space X is a set of pairs:

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

Definition 3.2. α level sets of a fuzzy set defined in X space are expressed as:

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

Definition 3.3. A TFN indicated by the fuzzy set is mathematically illustrated as:

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

Definition 3.4. $\widetilde{A} = (a_1, a_2, a_3)$ and $\widetilde{B} = (b_1, b_2, b_3)$ are two TFN s, and some arithmetical operations regarding fuzzy number s are as:

$$\begin{aligned} \text{Addition/Subtraction: } \widetilde{A} \pm \widetilde{B} &= (a_1 + b_1/a_1 - b_3, a_2 + b_2/a_2 - b_2, a_3 + b_3/a_3 - b_1) \\ \text{Multiplication: } \widetilde{A} \times \widetilde{B} &= (a_1 \cdot b_1, a_2 \cdot b_2, a_3 \cdot b_3) \\ \text{Division: } \widetilde{A}/\widetilde{B} &= (a_1/b_1, a_2/b_2, a_3/b_3). \end{aligned} \tag{4}$$

Definition 3.5 (Extension Principle)**.** Let F indicates fuzzy sets space (therefore $E \subseteq F$. So, for each arbitrary $m \in F, m : R \to [0,1]$, consider X to be the 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 respectively. 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 define a fuzzy set in Y.

$$\widetilde{A} = \left\{ \begin{array}{ll} \sup_{(x_1, x_2, \ldots, x_n) \in f^{-1}} \min\{m_1(x_1), m_2(x_2), \cdots, m_n(x_n)\} & f^{-1}(y) \neq 0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{5}$$

in which f^{-1} is the inverse image of f.

In many studies, Zadeh's Extension Principle introduced in Zadeh [37] is employed for the logarithmic transformation of possibilistic odds as fuzzy numbers representing observed output values.

Definition 3.6 (Possibilistic Odds)**.** Let $\mu_i, i = 1, \ldots, m$ be the possibility of success, $\mu_i = \text{poss}(Y_i \approx 1)$. The ratio, $\frac{\mu_i}{1-\mu_i} i = 1, \ldots, m$, is considered as "possibilistic odds" of the ith case, which defines "the possibility of success" relative to "the possibility of nonsuccess".

Logistic regression is used to identify the relationship between the probability of an output variable (*i.e.*, success) and input variables. Due to the vague status of cases relative to response categories, Bernoulli probability distribution cannot be assumed, as a result, probability of success ($P(Y_i = 1) = \pi_i$) cannot be calculated and modeled exactly based on explanatory variables. In this situation, the probability odds, $\left(\frac{\pi_i}{1-\pi_i}\right)$, the rate of success to fail, is irrelevant. To handle this situation, the possibility of success instead of the probability of success was introduced as in Definition 3.6 by Pourahmad *et al.* [16]. The cases are basically a linguistic variable defined by terms, $\mu_i \in \{\ldots, \text{ low, medium, high} \ldots\}$. These terms should be constructed as fuzzy numbers in such manner that the union of the supports covers the whole range of $[0, 1]$ as defined in Pourahmad *et al.* [9]. In this study, the output variable is represented by the linguistic term (None, Low, Medium, High or Critical). Adopting the same calculation steps given in [9,16], the output variable is converted into possibility of success. Zadeh's well-known extension principle is then employed for logarithmic (logit) transformation of possibility to "log of possibilistic odds" for linearity. Backwards transformation of $\widetilde{Y}_i$ to $\widetilde{\mu}_i$ is also possible, since the logit function is one-to-one and there is one and only one value for each element in the range. This logarithmic transformation can avert the nonlinear effects, so the model becomes intrinsically linear:

$$\widetilde{Y}_i = \ln \frac{\widetilde{\mu}_i}{1 - \widetilde{\mu}_i} = A_0 + A_1 \widetilde{x}_{i1} + \cdots + A_n \widetilde{x}_{\text{in}} \qquad i = 1, \ldots, m. \tag{6}$$

3.3. FLR

FLR is an integrated regression method that utilizes both the statistical regression analysis technique and the fundamentals of fuzzy set theory. FLR allows the dependent variable (output) to be fuzzy in binary, ordinal, and multinomial form, just as the independent variable/s (input) to be crisp (exact) or fuzzy in categorical or continuous form. As has been previously reported in the Literature Review, among the previous research in FLR, the works of Pourahmad *et al.* [9,16], Sohn *et al.* [21], Yoon and Choi [22], Yoon and Grzegorzewski [23], Namdari *et al.* [26], have comprehensively contributed to FLR. Especially, in Sohn *et al.* [21], Yoon and Choi [22], Yoon and Grzegorzewski [23] a fuzzy least-squares estimation method by defining new fuzzy operations was developed. The method provides us with regression parameters through one compact formula which was derived from the new fuzzy matrix and fuzzy matrix arithmetic operations defined in Sohn *et al.* [21], Yoon and Choi [22], Yoon and Grzegorzewski [23], which enabled to prove optimal or asymptotic properties easily. They also showed that the FLR method is actually a generalized fuzzy extension of statistical regression. They defined the fuzzy general multiple regression model as the following:

$$\widetilde{Y}_i = \beta_0 \oplus \beta_1 \widetilde{X}_{i1} \oplus \cdots \oplus \beta_p \widetilde{Y}_{ip} \oplus \Phi_i, \qquad i = 1, \ldots, n \tag{7}$$

where $\widetilde{X}_{ij}, \widetilde{Y}_i (j = 1, \ldots, m)$ are fuzzy numbers, and β_j are unknown regression crisp parameters to be estimated on the basis of fuzzy observations on $\widetilde{Y}_i$ and $\widetilde{X}_{ij}$. ϕ_i is assumed to be fuzzy error terms. The fuzzy numbers are $\widetilde{X}_{ij} = (l_{xij}, x_{ij}, r_{xij})$ and $\widetilde{Y}_{ij} = (l_{yi}, y_i, r_{yi})$, are represented by TFN, where l and r are the left and right endpoints calculated based on the left and right spread values respectively.

Diamond [20] proposed a metric in the set of all TFN. Let $F_T(\mathbb{R})$ denote the set of all TFN in $\mathbb{R}$. $\widetilde{X} = \langle v, \xi^l, \xi^r \rangle_\Delta, \widetilde{Y} = \langle w, \eta^l, \eta^r \rangle_\Delta$ are called TFN.

Definition 3.7. For $\widetilde{X}, \widetilde{Y} \in F_T(\mathbb{R})$, the distance between two fuzzy numbers, which is the error between the observed and the estimated fuzzy numbers, d, is defined as follows:

$$d^2\left(\widetilde{X}, \widetilde{Y}\right) = \left[w - \eta^l - \left(v - \xi^l\right)\right]^2 + \left[w + \eta^r - (v + \xi^r)\right]^2 + (w - v)^2. \tag{8}$$

Using metric d in model $\widetilde{Y}_i$, we obtain sums of squares error (SSE) as:

$$q(\beta_0,\beta_1,\ldots,\beta_M)=d^2\left(\widetilde{Y},\sum_{j=0}^{m}\beta_j\widetilde{X}_{ij}\right)=\left[l_{yi}-\sum_{j=0}^{m}\beta_j l_{xij}\right]^2+\left[q_i-\sum_{j=0}^{m}\beta_j p_{ij}\right]^2+\left[r_{yi}-\sum_{j=0}^{m}\beta_j r_{xij}\right]^2. \quad (9)$$

$\hat{\beta}_j$ can be obtained by minimizing the $Q=Q(\beta_0,\beta_1,\ldots,\beta_M)$. For each $k=0,1,\ldots,m$, $\frac{\partial Q}{\partial\hat{\beta}_j}=0$ results the normal equation, which has $\hat{\beta}_j$ as solutions,

$$\sum_{j=0}^{m}\hat{\beta}_j\sum_{i=1}^{n}(l_{xik}l_{xij}+p_{ik}p_{ij}+r_{xik}r_{xij})=\sum_{i=1}^{n}(l_{xik}l_{yi}+p_{ik}q_i+r_{xik}r_{yi}). \quad (10)$$

Here, we define the design matrix $\hat{X}$ as $[l_{xij},p_{ij},r_{xij}]_{n\times(m+1)}$, that is,

$$\widetilde{X}=\begin{bmatrix}(1,1,1)(l_{x11},p_{11}r_{x11}) & \cdots & (l_{x1p},p_{1m}r_{x1m})\\ \vdots & \ddots & \vdots\\ (1,1,1)(l_{x11},p_{11}r_{x11}) & \cdots & (l_{xnm},p_{nm}r_{xnm})\end{bmatrix}. \quad (11)$$

And define $\widetilde{y}$ as $[l_{yi},q_i,r_{yi}]_{n\times1}=[(l_{y1},q_1,r_{y1}),\ldots,(l_{yn},q_n,r_{yn})]^t$, then, the coefficient matrix of the system of normal equations. Consequently, we have a compact form as in Sohn *et al.* [21]:

$$\hat{\beta}=\left(\widetilde{X}^t\diamond\widetilde{X}\right)^{-1}\widetilde{X}^t\diamond\widetilde{y}. \quad (12)$$

Further explanations for the "$\diamond$" operator, and triangular fuzzy matrices $\widetilde{X}$ and $\widetilde{y}$ are presented in Yoon and Choi [22] and Yoon and Grzegorzewski [23]. In this study, an FLR model is fitted to estimate the CVSS base category using ICS data based on CVSS by applying the fuzzy least square estimation concept proposed by Yoon and Choi [22].

3.4. The performance criteria

We adopted MCI, Kim–Bishu Index (KB), MDM and Accuracy (ACC) to evaluate the performance of the model application.

Definition 3.8. MCI, proposed by Taheri and Kelkinnama [24], is the mean of capability index to measure the goodness-of-fit:

$$\mathrm{MCI}=\frac{1}{m}\sum_{i=1}^{m}I_{\mathrm{UI}}\left(\widetilde{w}_i,\widetilde{W}_i\right) \quad (13)$$

where Capability Index is defined by I_{UI}:

$$I_{\mathrm{UI}}=\frac{\mathrm{Card}(u\cap v)}{\mathrm{Card}(u\cup v)},\ \text{where Card}(u)=\begin{cases}\int_t u(t)\,\mathrm{d}t & \text{continuous case}\\ \sum_t u(t) & \text{discrete case.}\end{cases} \quad (14)$$

The "min" operator is used for the intersection of two fuzzy sets and the "max" operator for the union. It means that larger ratio of intersection to the union of the observed and the estimated fuzzy numbers indicates a better fit, and $0\le\mathrm{MCI}\le1$.

Definition 3.9. Kim and Bishu [28] defined the error of fitting by the ratio of the difference of membership values to the observed membership values. If the ratio of the total difference over the union of supports $\left(S_{\widetilde{Y}} \cup S_{\widehat{Y}}\right)$ to the observed membership values gets closer to zero, the two membership functions overlaps more, which corresponds a better goodness-of-fit. The KB is calculated as follows:

$$\text{Error} = \frac{\int_{S_{\widetilde{Y}} \cup S_{\widehat{\widetilde{Y}}}} \left|\widetilde{Y}(y) - \widehat{\widetilde{Y}}(y)\right| \mathrm{d}y}{\int_{S_{\widetilde{Y}}} \left|\widetilde{Y}(y)\right| \mathrm{d}y}. \tag{15}$$

Definition 3.10. MDM Criterion, introduced by Pourahmad *et al.* [16] as the mean of the membership values for the observed data evaluated in the estimated membership functions:

$$\text{MDM} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{W}_i(w_i) = \frac{1}{n} \exp\left(\left(\widetilde{W}_i\right) \frac{\mu_i}{1-\mu_i}\right) \tag{16}$$

MDM ≥ 0.5 provides a good fit.

Definition 3.11. ACC is the ratio of the number of correct classifications to the number of all classifications. In a multinomial classification, ACC is ratio of the sum of the values on the diagonal of the coincidence matrix (or confusion matrix) to the sum of all values in the matrix. The larger ACC means better performance in predicting the categories.

3.5. Metaheuristic Algorithms (MA)

MA are general algorithmic methods, usually nature-inspired, designed to approximate global optimum or solve complex problems in optimization by an efficient and guided search (typically based on stochastic operations not to be trapped in local optima), which explores a space of randomly generated or selected feasible solutions. Recently, MA are emerging as successful alternatives to more classical approaches also for solving optimization problems that include uncertain, stochastic, and dynamic information in their mathematical formulation [40]. There are many proposed MA available in the literature and they apply practically to a wide variety of problems. A general pseudo-code of any metaheuristic algorithm is outlined in Figure 12 [41]. Besides many attributes, MA are also classified in terms of their method in handling possible solutions; single solution approaches pivot a single candidate whereas population-based approaches keep multiple solutions to improve the objective function value. MA are typically applied to discrete combinatorial optimization problems, on the other hand there exist many studies carried out MA applications on continuous optimization problems.

Evolutionary Algorithm (EA), a population-based approach, applies biological mechanisms (*i.e.*, reproduction, mutation, recombination, and selection) and defines many operators such as mutations and crossover.

Simulated Annealing (SA), a single solution approach, is based on the process of melting and freezing metals on a molecular scale.

EA and SA are inarguably the two most extensively used nature-inspired algorithms [42]. In this study, EA and SA are used to improve the FLR model performance by optimizing the left-right spreads of TFN representing input variables.

3.6. Application

In this section, we conduct an FLR model with fuzzy input and fuzzy output data based on possibilistic odds defined by Pourahmad *et al.* [16], applying least-squares estimation method introduced by Yoon and Choi [22], using ICS vulnerability data. As noted in the previous section, there are seven input (independent) variables and one output (dependent) variable. Since all these categorical variables are in linguistic terms, which inherently possess uncertainty in their nature, they are represented by fuzzy numbers. A fuzzy number can be in the form of triangular, trapezoid or Gaussian, etc. In this study, TFN form is chosen to fuzzify the data for modeling. There

Pseudo-code for Generic Metaheuristic Algorithm

```
1: Procedure Metaheuristic
2:   Initialize population of candidate solutions
3:   Evaluate the initial solutions and remember the best one
4:   while (termination criteria not met) do
5:       generate new solutions by modifying existing solutions
6:       evaluate new solutions
7:        if new solutions are better than the existing solution then
8:              update population
9:       end if
10:             remember the best one
11:  end while
12:  report the best solution
16:  end procedure
```

FIGURE 12. The pseudo-code for Generic Metaheuristic algorithm.

are two reasons for this selection. First, since each level in variable should spread around a single value, such as "approximately 0.85", TFN can easily to capture this specific definition of CVSS metrics in the fuzzification step. If the data were represented by some particular intervals, trapezoidal fuzzy numbers would be choice of use to represent the variables in modeling [43]. In that case, the core of the trapezoid number actually represents an interval of numbers which are observed more frequently than the other numbers. Second, the arithmetic operations on triangular fuzzy matrices, to solve the normal equations for estimating the model parameters, have already been defined and successfully applied in many previous studies. A schematic representation of the stepwise model application is presented in Figure 13.

CVSS Specification Document available online [44] defines a numerical value of each metric level in the vulnerability data to calculate a continuous score using its specific equations. Actually, this gives a clear guidance to decide the approximate values of the center/modes and initial values of spread parameters of TFNs representing the regression variables. Taking the qualitative severity rating scale intervals defined in the Specification Document as basis, the output variable, which is the Base category (Critical, High, Medium, Low, and None), is fuzzified. The possibilistic odds are calculated and logit transformation is performed for FLR modeling (Tab. 3, Fig. 14). The support is totally covered by the TFNs representing different levels of corresponding output variable.

Likewise, applying the same steps, seven input variables are fuzzified. The endpoints of each TFN are calculated using the initial spread values acquired also from the CVSS Specification Document. Later, these initial spread value parameters are to be optimized by EA and SA application to improve the performance of the FLR model (Tab. 4, Figs. 15–19).

The fuzzified data ($n = 525$) was then randomly split into train ($n = 365$, 70%) and test set ($n = 160$, 30%). Another set of ICS vulnerability data (all newly released) is collected from the CISA database, to unbiasedly test/validate the final models with independent data (out-sample). An FLR model is fitted to the ICS vulnerability data based on CVSS, by applying the fuzzy least-square estimate approach used in Yoon and Choi [22]. The coefficients of the base model are estimated by FLR as follows:

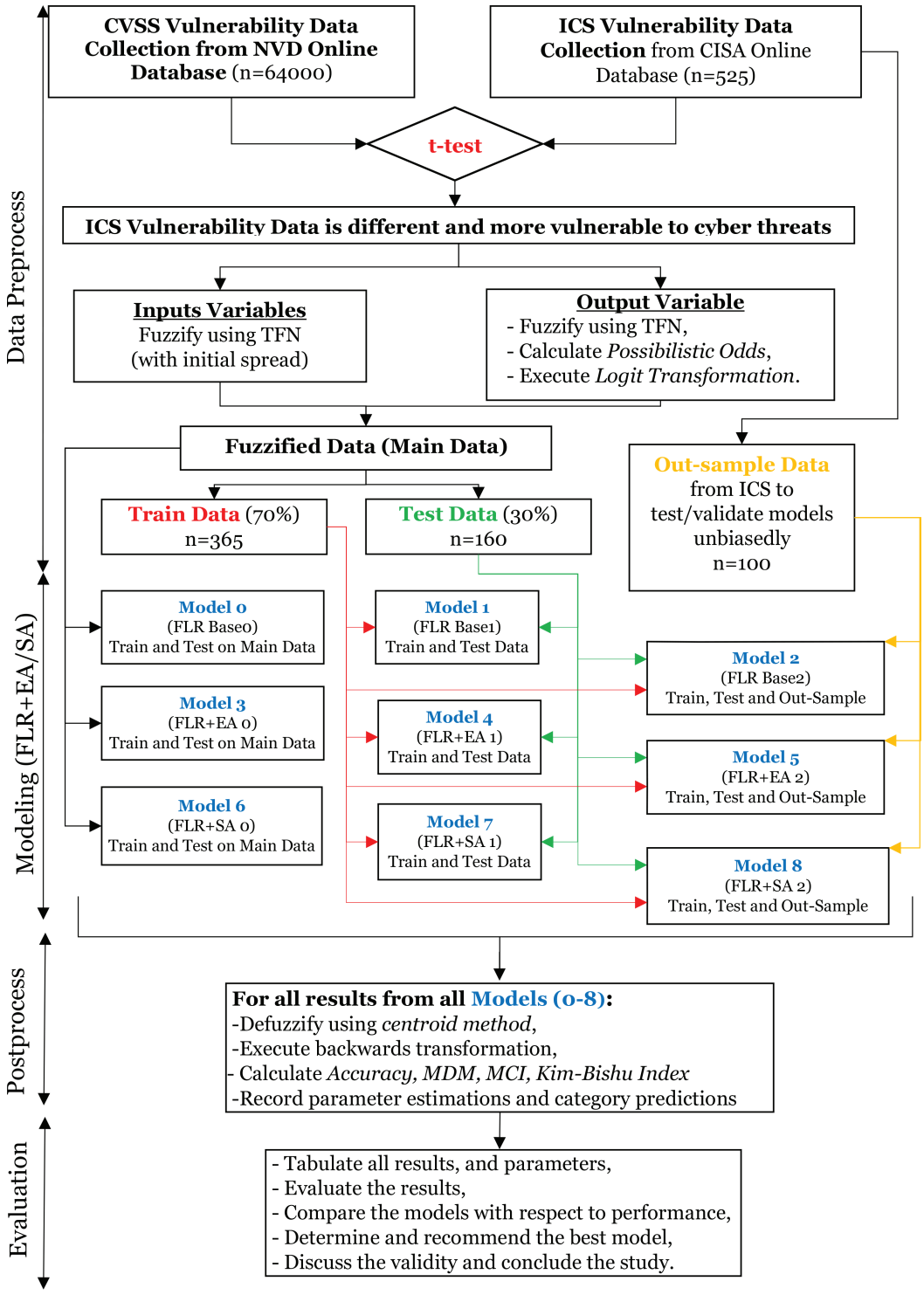

FIGURE 13. The stepwise procedure of the applied methodology.

TABLE 3. Fuzzification and transformation of output variable for FLR modeling.

Linguistic scales (category)	Triangular fuzzy numbers (endpoints)			$\widetilde{Y} = \ln\left(\frac{\mu}{1-\mu}\right)$		
	Left	Center	Right			
None	0.01	0.05	0.12	−4.5951	−2.9444	−1.9924
Low	0.08	0.245	0.42	−2.442	−1.1254	−0.3227
Medium	0.39	0.545	0.69	−0.447	0.1804	0.8001
High	0.6	0.795	0.92	0.4054	1.3553	2.44235
Critical	0.8	0.945	0.99	1.3862	2.8438	4.5951

TABLE 4. Fuzzification of input variables.

Input variables	Linguistic/verbal terms	Triangular fuzzy numbers		
		Left endpoint	Center	Right endpoint
Attack Vector (AV)	Network	0.70	0.85	0.99
	Adjacent Network	0.59	0.62	0.74
	Local	0.46	0.55	0.58
	Physical	0.01	0.20	0.45
Attack Complexity (AC)	Low	0.62	0.77	0.99
	High	0.010	0.44	0.6
Privileges Required (PR)	None	0.7	0.85	0.99
	Low (scope changed)	0.53	0.68	0.83
	Low (scope unchanged)	0.47	0.62	0.77
	High (scope changed)	0.35	0.5	0.65
	High (scope unchanged)	0.12	0.27	0.42
User Interaction (UI)	None	0.71	0.85	0.99
	Required	0.01	0.62	0.7
Confidentiality (C)	High	0.41	0.56	0.99
	Low	0.001	0.22	0.4
	None		0	
Integrity (I)	High	0.41	0.56	0.99
	Low	0.001	0.22	0.4
	None		0	
Availability (A)	High	0.41	0.56	0.99
	Low	0.001	0.22	0.4
	None		0	

$$\widetilde{Y} = \ln\left(\frac{\widetilde{\mu}}{1-\widetilde{\mu}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1\widetilde{X}_1 + \widehat{\beta}_2\widetilde{X}_2 + \cdots + \widehat{\beta}_7\widetilde{X}_7$$
$$= -3.2472 + 1.9690\ \widetilde{X}_1 + 0.3722\ \widetilde{X}_2 + 1.3599\ \widetilde{X}_3 + 0.6861\ \widetilde{X}_4 + 0.8064\ \widetilde{X}_5$$
$$+ 0.8404\ \widetilde{X}_6 + 1.2816\ \widetilde{X}_7.$$

Defuzzification (purification) using the centroid method and backwards transformations (for logit and possibilistic odds) are the postprocess steps of the model application.

3.7. Improvement of FLR model with EA and SA

In this section, we seek to improve the performance of the base FLR model. In the base model, the nonsymmetric spread values for fuzzy input variables are first determined using the guidance from CVSS Specifi-

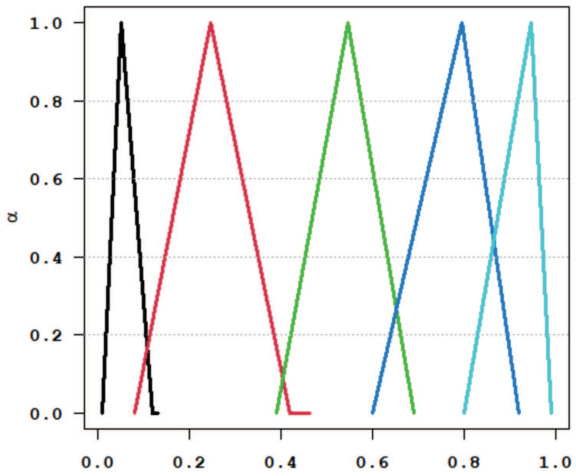

FIGURE 14. Fuzzy membership function of base category (None, Low, Medium, High, Critical).

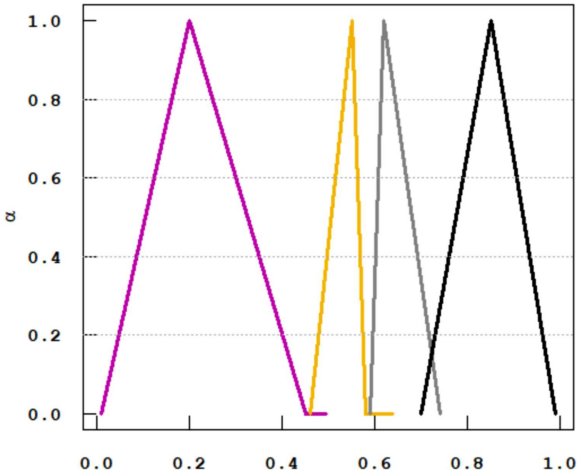

FIGURE 15. Attack Vector (Physical, Local, Adjacent, Network).

cation Document. A further question that arises here is whether tuning of non-symmetric spread values of the TFNs representing the input variables (no change in the value of the modes) can improve the performance of the FLR model in terms of accuracy. The base FLR model is reconstructed to have 14 non-symmetric spread parameters for seven input variables so that it can be optimized for better performance. The EA and SA were used for optimization separately.

EA algorithm is executed in MS Excel Solver Add-in [45] with the execution parameters given in Table 5. Along with these criteria, additional constraints are also defined, to ensure a reasonable spread over the support, left endpoints greater than zero, and right endpoint less than one.

As for SA application, Generalized Simulated Annealing (GenSA) package available in R statistical computing language is used [46]. GenSA has a built-in function that searches global minimum of a very complex non-linear objective function with a very large number of optima. The execution parameters for the application of SA are given in Table 6.

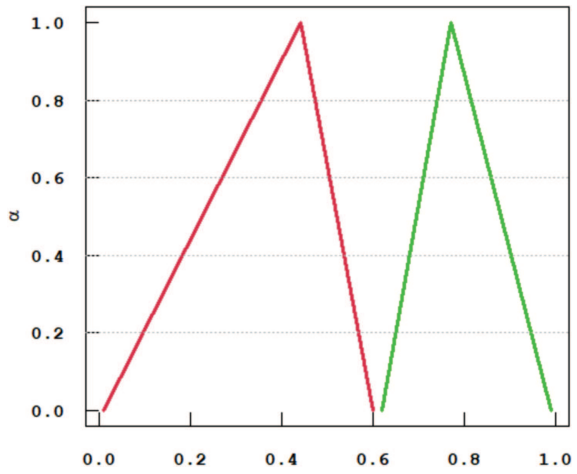

FIGURE 16. Attack Complexity (High, Low).

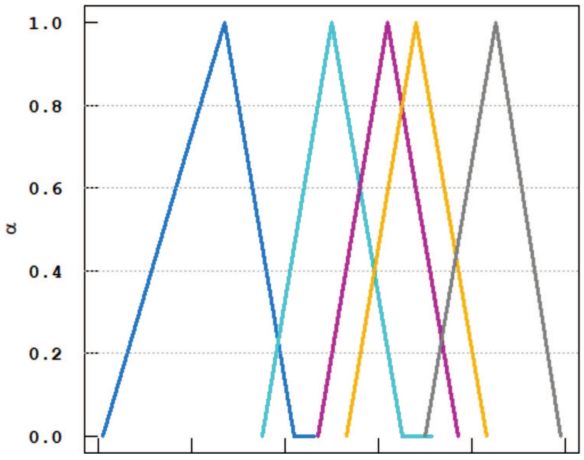

FIGURE 17. Privileges Required [High (SC, NSC), Low (SC, NSC), None].

TABLE 5. EA execution parameters.

EA execution parameters	Value set
Convergence	0.0001
Mutation rate	0.2
Population size	Default
Random seed	Random
Maximum time without improvement	100
Require bounds on variables	Yes
Max times (s)	Unlimited
Iteration limit	Unlimited

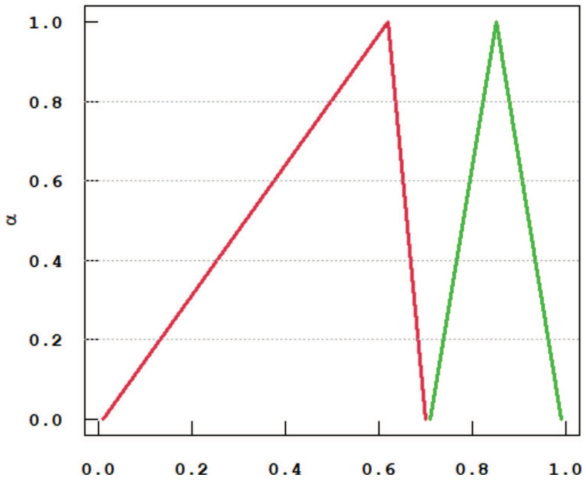

FIGURE 18. User Interaction (Required, None).

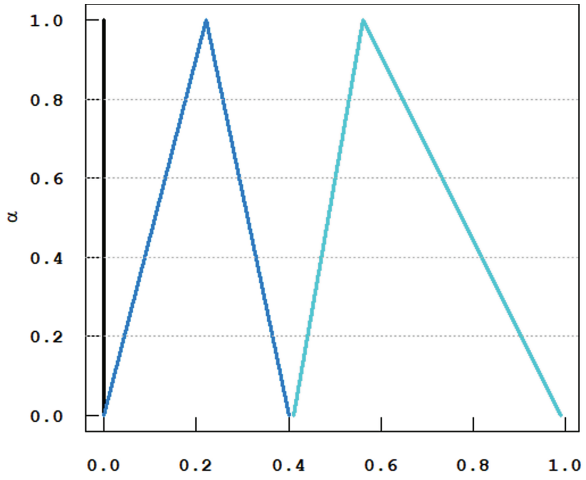

FIGURE 19. Confidentiality, Integrity and Availability (N, L, H).

TABLE 6. SA execution parameters.

SA execution parameters	Value set
Par (initial vector of values)	Random
Lower bounds	0.0001
Upper bounds	0.99
Max number of iterations	1000
Threshold.stop	Null
Nb.stop.improvement (no imp. max steps)	100
Smooth (differentiable obj. function	True
Temperature	400
Seed	Random

Table 7. Results.

Model	TRAIN	TEST	Out sample	Model	ACC	KB	MCI	MDM	INT	AV	AC	PR	UI	C	I	A
0	In-sample	In-sample		FLR	83%	0.88	0.47	0.62	−3.25	1.97	0.37	1.36	0.69	0.81	0.84	1.28
1	365	160		FLR	86%	0.78	0.52	0.69	−3.40	1.95	0.20	1.56	0.86	0.68	0.95	1.28
2	365		100	FLR	85%	0.86	0.50	0.71	−3.40	1.95	0.20	1.56	0.86	0.68	0.95	1.28
3	In-sample	In-sample		FLR+EA	91%	0.74	0.54	0.63	−6.53	2.14	1.56	2.32	2.25	1.33	1.34	1.74
4	365	160		FLR+EA	91%	0.68	0.57	0.70	−7.15	2.29	1.53	2.43	2.58	1.13	1.28	1.75
5	365		100	FLR+EA	90%	0.68	0.57	0.75	−7.15	2.29	1.53	2.43	2.58	1.13	1.28	1.75
6	In-sample	In-sample		FLR+SA	91%	0.77	0.53	0.67	−6.53	2.20	1.54	2.17	2.30	0.76	1.60	1.67
7	365	160		FLR+SA	88%	0.72	0.53	0.64	−6.58	2.21	1.82	2.20	2.51	0.88	1.72	1.41
8	365		100	FLR+SA	94%	0.68	0.56	0.61	−6.58	2.21	1.82	2.20	2.51	0.88	1.72	1.41

Table 8. The overall average results.

Model	Average Accuracy	Average Kim–Bishu Index	Average MCI	Average MDM
FLR BASE	0.85	0.84	0.50	0.67
FLR+EA	0.91	0.70	0.56	0.70
FLR+SA	0.91	0.72	0.54	0.64

4. Discussion

In this section, the performance of the model is evaluated and the results are discussed. There are in total nine different FLR models fitted, variations in the model are created by different data set usage (*i.e.*, in-sample/train/test) and existence of model improvement by EA/SA as presented in Figure 13. As for the performance criteria (Fig. 20), KB, MCI, MDM and ACC are used. The results obtained are outlined in Table 7 and overall average of performance is summarized in Table 8.

It can be readily observed that models with EA and SA have the same average ACC performance of 91% and they have approximately 6% of advantage in terms ACC over the Base models. Furthermore, the FLR models with optimized EA (Models 3–5) performed better on average with respect to three out of four criteria (KB, MCI, MDM). When we examine the results, we can claim three things. First, when the spread values of TFNs are tuned by EA/SA, the performance of FLR model improves. Second, in terms of improvement the performance, EA is effectively superior to SA. Third, when we test the models with the out-sample data, the performance values are in a reasonable range with no outliers. This shows the prediction stability of the application. It is also assumed to be a good indication that even if there are different designs for modeling, the values of estimated parameters are close to each other within groups, so it can be inferred that the model is insensitive to both the way of sampling data for modelling and different approaches in developing FLR models (Fig. 21). In-sample modeling, in a sense, should achieve the best performance since train and test data are the same, but this is not the case. The reason why that is the initial spread is defined based on the CVSS Specification Document, when these initial values are optimized they yield better results.

When we examine the coincidence matrices (Tabs. 9–17), the predictions are mostly on the diagonals of the matrices, which indicates the ACC of the models. The correct classifications are shown in bold font. No bad/unacceptable misclassifications occurred, since most of the misclassifications, higlighted in red color, are dispersed in the "one-step-away" neighborhood of the diagonal. Predominantly, a reasonable number of "CRITICAL" and "MEDIUM" observations are misclassified as "HIGH". An inclusion of a penalty matrix in EA/SA optimization formulation may decrease the number of such misclassifications.

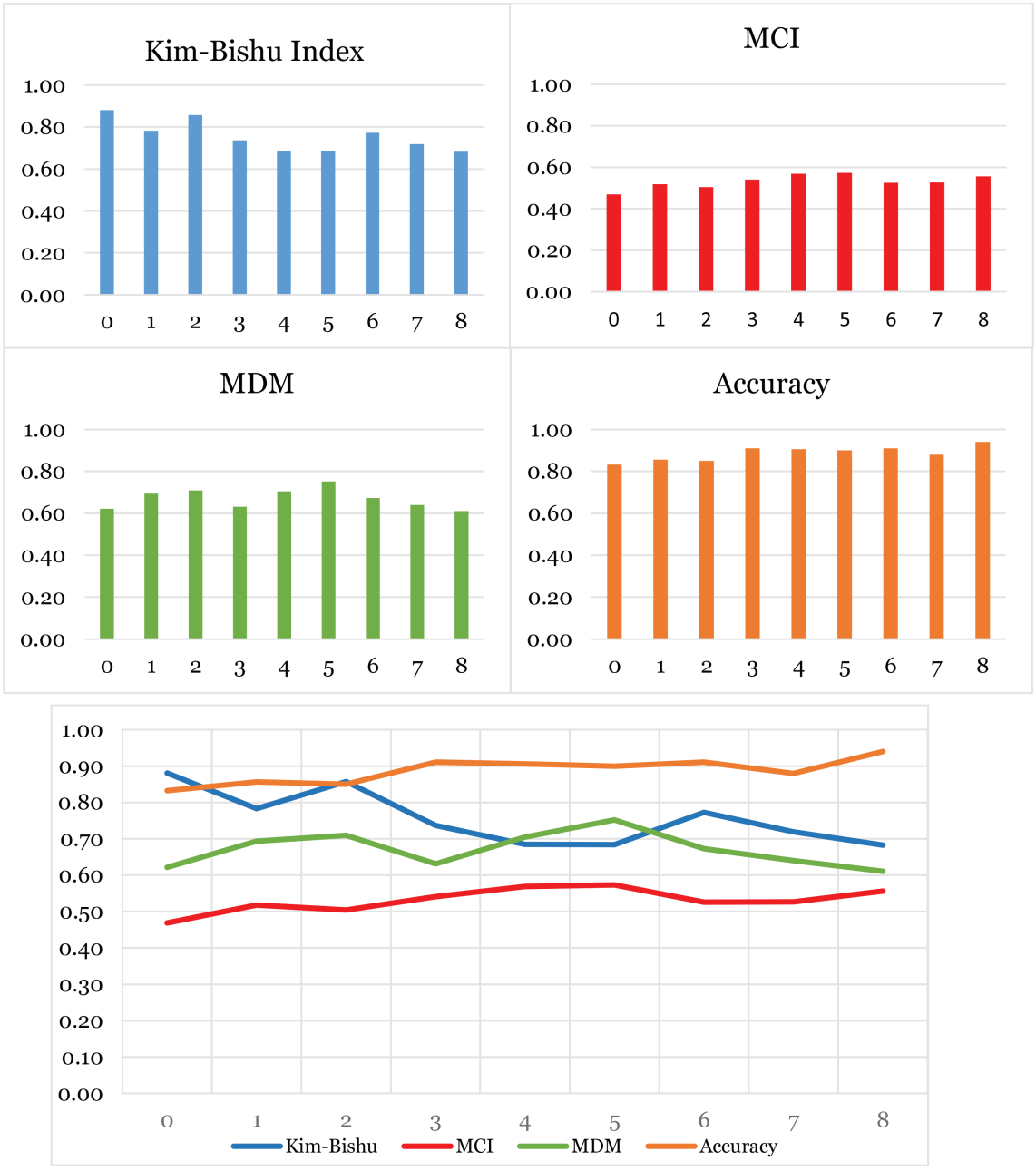

FIGURE 20. The values of the performance criteria.

TABLE 9. The coincidence matrix for Model 0 (base FLR with in-sample data).

Model 0		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**89**	9			98
	HIGH	30	**226**	23		279
	MEDIUM	1	10	**118**	14	143
	LOW			1	**4**	5
	Total	120	245	142	18	525

TABLE 10. The coincidence matrix for Model 1 (base FLR with 365 Train and 160 test set).

Model 1		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**21**	2			23
	HIGH	9	**85**	6		100
	MEDIUM		2	**29**	3	34
	LOW			1	**2**	3
	Total	30	89	36	5	160

TABLE 11. The coincidence matrix for Model 2 (base FLR with 365 Train and 100 out-test set).

Model 2		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**14**	1			15
	HIGH	2	**51**	8		61
	MEDIUM		2	**20**	2	24
	LOW			0	**0**	0
	Total	16	54	28	2	100

TABLE 12. The coincidence matrix for Model 3 (FLR+EA with in-sample data).

Model 3		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**95**				95
	HIGH	24	**234**	6		264
	MEDIUM	1	11	**133**	2	147
	LOW			3	**16**	19
	Total	120	245	142	18	525

TABLE 13. The coincidence matrix for Model 4 (FLR+EA with 365 Train and 160 test set).

Model 4		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**23**				23
	HIGH	7	**86**	2		95
	MEDIUM		3	**32**	1	36
	LOW			2	**4**	6
	Total	30	89	36	5	160

TABLE 14. The coincidence matrix for Model 5 (base FLR with 365 Train and 100 out-test set).

Model 5		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**14**				14
	HIGH	2	**52**	2		56
	MEDIUM		2	**24**	2	28
	LOW			2		2
	Total	16	54	28	2	100

TABLE 15. The coincidence matrix for Model 6 (FLR+SA with in-sample data).

Model 6		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**95**				95
	HIGH	24	**236**	11		271
	MEDIUM	1	9	**128**	4	142
	LOW			3	**14**	17
	Total	120	245	142	18	525

TABLE 16. The coincidence matrix for Model 7 (FLR+SA with 365 Train and 160 test set).

Model 7		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**23**				23
	HIGH	7	**85**	5		97
	MEDIUM		4	**30**	3	37
	LOW			1	**2**	3
	Total	30	89	36	5	160

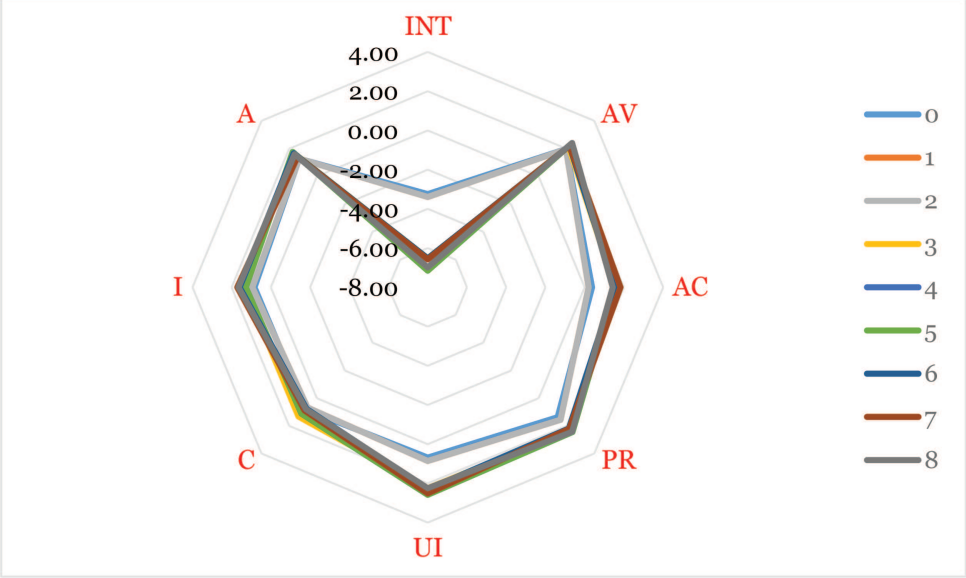

FIGURE 21. The estimated parameters for all models.

TABLE 17. The coincidence matrix for Model 8 (FLR+SA with 365 Train and 100 out-test set).

Model 8		OBSERVED				
		CRITICAL	HIGH	MEDIUM	LOW	Total
PREDICTED	CRITICAL	**14**				14
	HIGH	2	**53**			55
	MEDIUM		1	**27**	2	30
	LOW			1	**0**	1
	Total	16	54	28	2	100

Another interesting point to discuss is the pattern/shape of MF after EA/SA optimizes the spread parameters. In a sense, we find, if not optimal, an approximate MF by means of tuning the spread values. After the optimization process, the final MF turn out not to cover the support on the interval of $[0, 1]$ (Fig. 22). This is the same essence of "narrow fuzzy number" and "expanded fuzzy number" previously explored in Sohn *et al.* [21]. An expanded fuzzy number covers the whole support, there are overlaps in the lower α-cut sets, and on the other hand a narrow fuzzy number does not. In this study, an MF, which is defined as an expanded fuzzy number at the beginning of modeling, transformed into a narrow fuzzy number. The underlying reason for this change in the shape of MF after optimization is that this variable is actually "less fuzzy" than initially assumed. This may also provide an approximate insight to decide, whether or not the type of fuzzy number (and its spread parameters) is an effective representative of that variable.

No overlapping and whole support is not covered.

5. CONCLUSION

In real life, it is frequently more practical to define qualifications or categories using linguistic/verbal terms (such as low, medium, and high) rather than exact numbers. Similarly, in many researches based on human

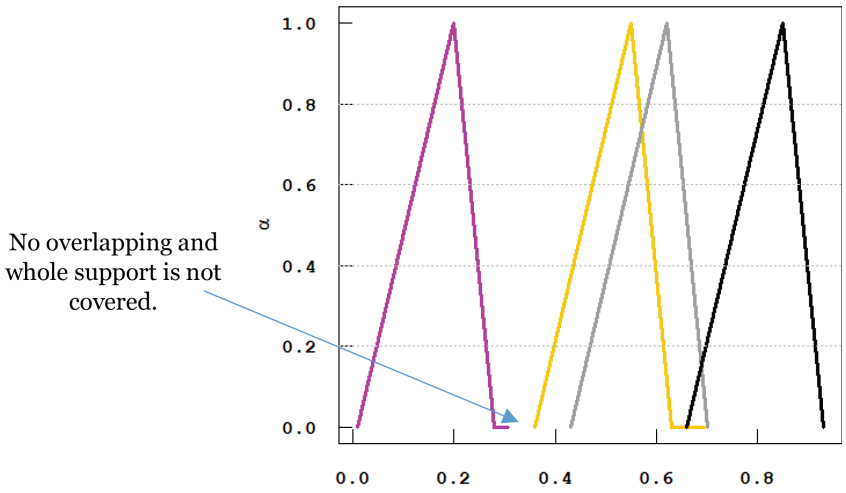

FIGURE 22. The MF of AV after application of EA in Model 3.

assessment, using linguistic terms to define categories is reasonably more applicable to describe data for modeling. Since human assessment innately bears uncertainty, the exact numerical boundaries of linguistic terms are imprecise and a probability distribution cannot be assumed for the output variable. In some cases, the basic statistical assumptions for classical statistical models are not fulfilled. The vulnerability data for ICS based on CVSS mostly conforms to the definition given above, it is indeed defined by some linguistic categorical terms and the values of variables are assigned by subjective expert judgment.

Such cases, although there are many other alternative methods available, fuzzy regression models based on fuzzy set theory introduced by Zadeh [37] can be applied since it is one of the an appropriate alternative methods to analyze categorical data with uncertainty and vagueness. Fuzzy regression models can efficiently contain imprecise information and systemically analyze it by transforming the data into fuzzy numbers. The considerations given above motivated us to apply fuzzy regression model to build a prediction model specifically focuses on ICS vulnerability data.

In this study, to illustrate the applicability, we fitted an FLR model to ICS vulnerability data. Furthermore, the model is effectively improved by applying metaheuristic algorithms for optimizing the spread of fuzzy numbers representing input variables. The model is assessed using ACC, KB, MCI and MDM performance criteria. The model achieved 91% of ACC. The findings this study show that FLR, as an alternative method, using both fuzzy input and fuzzy multinomial output data, can be successfully applicable to ICS vulnerability data based on CVSS.

The application of metaheuristic algorithms to optimize the spread values of the fuzzy numbers corresponding to input variables notably improved the models performance. After the application of metaheuristic algorithms, the shape of the MFs also changed. Some of the supports of the "new MFs" became narrower, indicating "how fuzzy the variable really is".

For further studies, this model can be expanded using trapezoid fuzzy numbers, different defuzzification methods may be explored, a penalty matrix can be defined in coding the metaheuristic algorithms so that bad misclassifications can be prevented, one of the other alternative/classical methods can be used for modeling to compare the results with the ones from FLR model, the goal programming approach may be applied in the optimization phase to balance a trade-off in conflicting objectives (such as ACC *vs.* KB).

Acknowledgements. The authors would like to thank the editors and reviewers for their constructive suggestions and corrections to enhance the clarity and the quality of this article.

References

[1] IBM, What is Industry 4.0? IBM. https://www.ibm.com/topics/industry-4-0 (accessed 21.12.2021, 2021).

[2] U.S. Department of Commerce, Information Security. [Online] Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf (2012).

[3] U.S. Department of Commerce, Computer Security. [Online] Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30.pdf (2002).

[4] National Institute of Standards and Technology (NIST), Vulnerabilities. https://nvd.nist.gov/vuln (accessed 19.06.2021, 2021).

[5] D. Braue, Global Cybersecurity Spending to Exceed $1.75 Trillion from 2021 to 2025. https://cybersecurityventures.com/cybersecurity-spending-2021-2025/ (accessed 21.12.2021, 2021).

[6] U.S. Department of Commerce, National Institute of Standards and Technology (NIST), National Vulnerability Database. NIST. https://nvd.nist.gov/vuln/data-feeds (accessed 23.12.2021, 2021).

[7] U.S. Department of Commerce, National Institute of Standards and Technology (NIST), CVSS Vulnerability Metrics. https://nvd.nist.gov/vuln-metrics/cvss (accessed 22.12.2021, 2021).

[8] Cybersecurity & Infrastructure Security Agency, ICS-CERT Advisories. US Department of Homeland Security. https://www.cisa.gov/uscert/ics/advisories (accessed 23.12.2021, 2021).

[9] S. Pourahmad, S.M.T. Ayatollahi, S.M. Taheri and Z.H. Agahi, Fuzzy logistic regression based on the least squares approach with application in clinical studies. *Comput. Math. App.* **62** (2011) 3353–3365.

[10] N. Chukhrova and A. Johannssen, Fuzzy regression analysis: systematic review and bibliography. *Appl. Soft Comput.* **84** (2019) 105708.

[11] FIRST.org, Common Vulnerability Scoring System version 3.1: Specification Document. https://www.first.org/cvss/specification-document (accessed 22.12.2021, 2021).

[12] M.G. Dondo, A vulnerability prioritization system using a fuzzy risk analysis approach. in Proceedings of the IFIP Tc 11 23rd International Information Security Conference. Springer US, Boston, MA (2008) 525–540.

[13] I.V. Anikin, Using fuzzy logic for vulnerability assessment in telecommunication network, in 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). IEEE (2017) 1–4. DOI: 10.1109/ICIEAM.2017.8076444.

[14] K. Gencer and F. Başçiftçi, The fuzzy common vulnerability scoring system (F-CVSS) based on a least squares approach with fuzzy logistic regression. *Egypt. Inf. J.* **22** (2021) 145–153.

[15] Y. Gao and Q. Lu, A fuzzy logistic regression model based on the least squares estimation. *Comput. Appl. Math.* **37** (2018) 3562–3579.

[16] S. Pourahmad, S.M.T. Ayatollahi and S.M. Taheri, Fuzzy logistic regression: a new possibilistic model and its application in clinical vague status. *Iran. J. Fuzzy Syst.* **8** (2011) 1–17.

[17] S. Mustafa, S. Asghar and M. Hanif, Fuzzy logistic regression based on least square approach and trapezoidal membership function. *Iran. J. Fuzzy Syst.* **15** (2018) 97–106.

[18] H. Tanaka, S. Uejima and K. Asai, Linear regression analysis with fuzzy model. *IEEE Trans. Syst. Man Cybern.* **12** (1982) 903–907.

[19] N.Y. Pehlivan and A. Yonar, An integrated approach for fuzzy logistic regression. *Istatistikçiler Dergisi: Istatistik ve Aktüerya* **11** (2018) 42–54.

[20] P. Diamond, Fuzzy least squares. *Inf. Sci.* **46** (1988) 141–157.

[21] S.Y. Sohn, D.H. Kim and J.H. Yoon, Technology credit scoring model with fuzzy logistic regression. *Appl. Soft Comput.* **43** (2016) 150–158.

[22] J.H. Yoon and S.H. Choi, Fuzzy least squares estimation with new fuzzy operations, in Synergies of Soft Computing and Statistics for Intelligent Data Analysis, edited by R. Kruse, M.R. Berthold, C. Moewes, M.Á. Gil, P. Grzegorzewski and O. Hryniewicz. Springer Berlin Heidelberg, Berlin, Heidelberg (2013) 193–202.

[23] J.H. Yoon and P. Grzegorzewski, On optimal and asymptotic properties of a fuzzy L2 estimator. *Mathematics* **8** (2020) 1956.

[24] S.M. Taheri and M. Kelkinnama, Fuzzy least absolutes regression, in 2008 4th International IEEE Conference Intelligent Systems. Vol. 2. IEEE (2008). DOI: 10.1109/IS.2008.4670509.

[25] G. Atalik and S. Senturk, A new approach for parameter estimation in fuzzy logistic regression. *Iran. J. Fuzzy Syst.* **15** (2018) 91–102.

[26] M. Namdari, A. Abadi, S.M. Taheri, M. Rezaei, N. Kalantari and N. Omidvar, Effect of folic acid on appetite in children: ordinal logistic and fuzzy logistic regressions. *Nutrition* **30** (2014) 274–278.

[27] M. Namdari, J. Yoon, A. Abadi, S.M. Taheri and S. Choi, Fuzzy logistic regression with least absolute deviations estimators. *Soft Comput.* **19** (2015) 909–917.

[28] B. Kim and R.R. Bishu, Evaluation of fuzzy linear regression models by comparing membership functions. *Fuzzy Sets Syst.* **100** (1998) 343–352.

[29] F. Salmani, S.M. Taheri and A. Abadi, A forward variable selection method for fuzzy logistic regression. *Int. J. Fuzzy Syst.* **21** (2019) 1259–1269.

[30] R. Xu and C. Li, Multidimensional least-squares fitting with a fuzzy model (in English). *Fuzzy Sets Syst.* **119** (2001) 215–223.

[31] F. Salmani, S.M. Taheri, J.H. Yoon, A. Abadi, H. Alavi Majd and A. Abbaszadeh, Logistic regression for fuzzy covariates: modeling, inference, and applications. *Int. J. Fuzzy Syst.* **19** (2017) 1635–1644.

[32] R. Nikbakht and A. Bahrampour, Determining factors influencing survival of breast cancer by fuzzy logistic regression model (in English). *J. Res. Med. Sci.* **22** (2017) 135–135.
[33] A. Behnampour, A. Biglarian and E. Bakhshi, Application of fuzzy logistic regression in modeling the severity of autism spectrum disorder (in English). *Jorjani Biomed. J.* **7** (2019) 49–60.
[34] M. Kelkinnama and S.M. Taheri, Fuzzy least-absolutes regression using shape preserving operations. *Inf. Sci.* **214** (2012) 105–120.
[35] T.M.B. Bennaser, *Fuzzy logistic regression for detecting differential DNA methylation regions*, Ph.D. in Applied Mathematics Doctoral Dissertations, Mathematics and Statistics, Missouri University of Science and Technology, USA (2020).
[36] W. Anggraeni, S. Sumpeno, E.M. Yuniarno, R.F. Rachmadi, A.B. Gumelar and M.H. Purnomo, Prediction of dengue fever outbreak based on climate factors using fuzzy-logistic regression, in 2020 International Seminar on Intelligent Technology and its Applications (ISITIA), 22–23 July 2020. IEEE (2020) 199–204. DOI: 10.1109/ISITIA49792.2020.9163708.
[37] L.A. Zadeh, Fuzzy sets. *Inf. Control* **8** (1965) 338–353.
[38] E. Çeven and Ö. Özdemir, Using fuzzy logic to evaluate and predict Chenille Yarn's shrinkage behaviour. *Fibres Text. Eastern Eur.* **15** (2007) 55–59.
[39] A.H. Gandomi, X.-S. Yang, S. Talatahari and A. Alavi, Metaheuristic algorithms in modeling and optimization, in Metaheuristic Applications in Structures and Infrastructures. Elsevier, London (2013) 1–24.
[40] L. Bianchi, M. Dorigo, L.M. Gambardella and W.J. Gutjahr, A survey on metaheuristics for stochastic combinatorial optimization. *Nat. Comput.* **8** (2009) 239–287.
[41] F.A. Hashim, K. Hussain, E.H. Houssein, M.S. Mabrouk and W. Al-Atabany, Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. *Appl. Intell.* **51** (2021) 1531–1551.
[42] H. Rajabi Moshtaghi, A. Toloie Eshlaghy and M.R. Motadel, A comprehensive review on meta-heuristic algorithms and their classification with novel approach (in English). *J. Appl. Res. Ind. Eng.* **8** (2021) 63–89.
[43] M. Voskoglou, Use of the triangular fuzzy numbers for student assessment. *Am. J. Appl. Math. Stat.* Preprint arXiv:1507.03257 (2015).
[44] FIRST.org, Common Vulnerability Scoring System v3.0: Specification Document. FIRST https://www.first.org/cvss/v3.0/specification-document#n3 (accessed 26.12.2021, 2021).
[45] F. Solvers. Excel solver – change options for evolutionary solving method. Frontline Solvers. https://www.solver.com/excel-solver-change-options-evolutionary-solving-method (accessed 27.12.2021, 2021).
[46] CRAN, Package "GenSA". https://cran.r-project.org/web/packages/GenSA/GenSA.pdf (accessed 27.12.2021, 2021).