Risk Assessment Through Big Data: An Autonomous Fuzzy Decision Support System

Mohammad Siami, Mohsen Naderpour, Fahimeh Ramezani, Jie Lu, *Fellow, IEEE*

***Abstract*— Decision support systems (DSSs) are computer-based systems that support managers in making operational, tactical, and strategic decisions. DSSs have been built to assist with a wide range of applications; however, in this paper, we are primarily concerned with risk assessment, which helps decision-makers to evaluate the risk of events. In recent years, advancements in big data processing, artificial intelligence, and machine learning have provided new opportunities for businesses to use these technologies for risk assessment. Yet, using these techniques with massive unlabeled data in an uncertain situation is challenging. This paper presents an autonomous fuzzy decision support system (AFDSS) for risk assessment that uses advanced artificial intelligence, unsupervised learning, and fuzzy logic. The model learns from big data characterized by uncertainty and a lack of labels for maximum utility. In an innovative approach, fuzzy clustering first extracts events from big data. Then, the risks associated with those events are assessed via a fuzzy inference system. New events are subsequently predicted based on their similarity to previously evaluated events. Evaluations of AFDSS with a real-world insurance dataset containing 500,000 journeys by 2500 drivers show that the proposed model can consistently assess risk in the big data environment. These results were drawn from a sensitivity analysis where all input parameters were changed using optimistic, pessimistic, and neural strategies. Performance was good across all three categories.**

Index Terms— Big data, risk assessment, decision making, fuzzy systems

I. Introduction

Decision support systems (DSS) are computerized systems that help decision makers find the best options from a range of alternatives by learning and analyzing historical and current data [1]. Multiple types of DSSs, such as model-driven, data-driven, and knowledge-driven systems, have been widely applied in different domains and have become more prevalent in recent years [2]. These systems require a number of criteria, a set of options, and a ranking methodology to sort alternatives according to the goal of a decision-maker. Various criteria need to be defined; experts and stakeholders need to answer related questions; numerical values need to be processed; and one option needs to be selected [3]. Therefore, they can be very costly and time-consuming. However, in recent years, DSS modeling has changed significantly. Technological improvements in database engineering, information technology, and the Internet of Things (IoT) have increased data volume, variety, and velocity [4]. Further, recent advancements in artificial intelligence and advanced analytical techniques are providing new opportunities for decision-makers to create unimaginable value from big data [5, 6].

Risk assessment is the process of evaluating the possibility of dangerous events and the activities involved in them [7]. It involves a broad range of activities to assess the probability and severity of an accident using quantitative or qualitative approaches [8]. Calculating risk exposure in big data with a complicated data structure is a challenging problem requiring a complex process. Identifying quantitative and qualitative criteria is the primary concern for a risk assessment DSS [9]. Over the past few years, machine learning and artificial intelligence algorithms have provided new opportunities for business experts to automate the risk assessment process in uncertain situations using advanced analytic techniques [10]. For example, neural networks perform outstandingly well at health risk assessment [11]. Random forest classifiers are doing well at financial risk assessment [12], and boosting algorithms work efficiently for fraud risk assessment [13]. Any of these models might be the right choice for risk assessment in a big data environment – but only when labeled data is available.

Therefore, this study's main motivation is to build a risk assessment framework that can handle unlabeled data challenges in a big data environment. Labeled data is a critical resource for any machine learning project, and a lack of labeled data can raise many challenges for these kinds of projects. Thus, we introduce a new autonomous fuzzy decision support system (AFDSS) to assess risk in big data environments characterized by uncertainty and a lack of labeling. As such, unsupervised learning algorithms and fuzzy logic are integral to the framework. The unsupervised algorithms handle the unlabeled data using degrees of truth instead of crisp logic. The fuzzy logic simulates the human mind in a complex risk decision making problem to reduce mathematical challenges.

The paper addresses the following research questions:

1) The availability of labeled data is critical to many machine learning algorithms, and these models typically minimize their cost functions based on access to labeled data. Thus, how can we build an autonomous decision support system that learns from unlabeled big data using advanced artificial intelligence and machine learning algorithms to support risk assessment in complex situations?

The authors are with Australian Artificial Intelligence Institute (AAII), Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), NSW 2007, Australia (e-mail: Mohammad.Siaminamini@uts.edu.au; Mohsen.Naderpour@uts.edu.au; Fahimeh.Ramezani@uts.edu.au; Jie.Lu@uts.edu.au).

2) Emerging technologies such as IoT, social media, and sensor technologies generate a massive amount of data with complex structures. The data generated by these devices are noisy, inconsistent, and incomplete [14]. Therefore, how can we assess the risk of IoT generated time-series data in uncertain situations with high precision?
3) How can we evaluate the effectiveness of the proposed decision support system using a case study with appropriate measures and performance indicators?

To answer these questions, we propose a novel decision support system that, using big data, learns to extract various criteria for risk assessment via an unsupervised learning algorithm. Within the framework, a fuzzy inference system evaluates the risks, and the system's performance is evaluated in the context of car insurance. Therefore, the main contributions of this paper are as follows:

- First, an autonomous fuzzy decision support system that learns from big data is proposed. The proposed framework has five major components: 1) data preparation; 2) risk factor mining; 3) fuzzy risk modeling; 4) event detection; and 5) risk calculation. Although the framework is designed to handle risk assessment, it is still general and can also be proposed toward other multi-criteria decision making.
- Second, an unsupervised learning framework is proposed to automatically extract different risk factors for decision making. A lack of labeled data is a fundamental challenge for all machine learning and artificial intelligence algorithms. Thus, in this study, we deal with the current challenges of unlabeled data in the big data domain, which remains an open problem. The provided algorithm is a two-step clustering algorithm incorporating a self-organizing map (SOM) and fuzzy clustering. The SOM reduces the complexity of the data, and the fuzzy clustering categorizes the input dataset.
- Third, a new fuzzy decision support system is proposed to handle uncertainty and a lack of confidence in the noisy data. The big data generated by sensor technologies contains noise, and analyzing them can impair analytical results. Therefore, we use the capability of fuzzy inference systems to decrease uncertainty and a lack of confidence in this domain.
- Finally, we undertook a case study to demonstrate and evaluate the proposed system in a practical sense. Using big data collected by smartphones, we assess the risk of car journeys for usage-based insurance, evaluating the results using sensitivity analysis to show confidence.

The rest of this paper is organized as follows. Section II provides a literature review. Section III presents the proposed model. Section IV explains the risk assessment case study, and validation results are seen in Section V. Section VI provides a discussion, and Section VII concludes the paper and describes future works.

II. Literature Review

The literature review serves as a foundational exploration of our Autonomous Fuzzy Decision Support System (AFDSS), providing a comprehensive overview of the methodologies and concepts that underpin its development. It directs focus towards pivotal components such as the self-organizing map (SOM), which enables unsupervised learning from vast datasets while concurrently streamlining complexity. Additionally, attention is drawn to the significance of fuzzy C-mean clustering in facilitating the extraction of actionable insights for risk-based decision-making within a fuzzy framework.

A. Self-organizing map

A SOM is a colloquial unsupervised learning algorithm that generates a map of the neurons in a neural network from input records. It has many applications, such as vector quantization, dimension reduction, and data visualization [15]. Decreasing computation costs is the primary advantage of a SOM because most clustering algorithms are based on calculating the distance between records to categorize data in similar groups. Therefore, clustering algorithms have very high computational costs, even with only small volumes of data. A SOM decreases the complexity of data by abstracting the input data to several prototypes. Then a clustering algorithm is used to cluster the prototypes instead of the full dataset [16].

Moreover, a SOM algorithm is not sensitive to noisy data because each node represents a group of records; thus, noise is less likely to affect it [17]. By contrast, detecting outliers is one of the most significant weaknesses of a SOM. There are generally only a few outlying data points, and SOMs have difficulty generating a suitable prototype to represent that data [18].

SOM algorithms map the input data into a two-dimensional map with N nodes that is either a rectangular or hexagonal grid. Each node has d features, which is equal to the number of input features with a weight $\omega_i = [\omega_{i1}, \omega_{i2}, \dots, \omega_{id}]^T$. The algorithm is iterative. In step t, a data sample $x(t)$ is selected randomly from the training data, and the most similar node to $x(t)$ is selected according to the calculated distances between $x(t)$ and all the nodes with the following:

$$c = argmin(dist(x(t), \omega_i), \quad \forall\, i\ in\ [1,2, \dots, N] \qquad (1)$$

where $dist(x(t), \omega_i)$ is equal to the distance of the sample x(t) with the i-th node.

Then, the winning neuron updates both itself and its neighboring neurons using the following rule:

$$\omega_k(t+1) = \begin{cases} \omega_k(t) + \gamma(t) h_{kc}(t).\left(x(t) - \omega_j(t)\right), \forall k \in N_c \\ \omega_k(t), \qquad else \end{cases} \qquad (2)$$

where N_c is the winning neuron's neighbors, and $\gamma(t)$ is the learning rate, which reduces in each iteration (t) according to the following equation:

$$\gamma(t) = \gamma_0 . exp\left(-\alpha . \frac{t}{\tau}\right) \qquad (3)$$

Here, γ_0 is the initial learning rate, α is the exponential decaying constant, and τ is the maximum number of iterations. $h_{kc}(t)$ is a

neighborhood kernel function that indicates the distance of the k[th] neuron to the winning neuron c, as calculated by:

$$h_{kc} = \exp\left(-\frac{[(x_k - x_c)^2 + (y_k - y_c)^2]}{2(\sigma(t)^2)}\right) \quad (4)$$

where $\sigma(t)$ is equal to the width of the neighborhood function. This decreases in each iteration *t* by

$$\sigma(t) = \gamma_0 . \exp\left(-\frac{t}{\tau} . \log(\sigma_0)\right) \quad (5)$$

where σ_0 is the initial width [19].

B. Fuzzy C-means (FCM) clustering

Fuzzy and k-means clustering algorithms are very similar and have shown outstanding performance at pattern recognition. K-means clustering provides a discrete clustering result, which means each member is part of only one cluster. Conversely, fuzzy clustering offers more information than k-means by providing a range score between zero and one, which is the similarity between members and clusters [20]. This characteristic is very useful for recognizing patterns in driving styles. Interestingly, most driving behaviors are similar, making it hard for transportation experts to easily specify a discrete cluster for each driving behavior. However, using fuzzy clustering for pattern recognition can detect unique and particular driving styles, as follows.

Let $X = \{X_1, X_2, \dots, X_n\}$ and be a multi-dimensional input dataset. A fuzzy clustering process categorizes n items into c clusters by developing an optimization process with the following objective function [21]:

$$J_m^{(FCM)}(U, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \left\| x_j - v_i \right\|_2^2 \quad (6)$$

where u_{ij} is a membership function, $\forall j = 1, \dots, n, \; and \; \sum_{i=1}^{c} u_{ij} = 1$. $V = \{v_1, v_2, \dots, v_c\}$ and represents the center of the clusters. m is a fuzzy factor that should be > 1 (usually *m* is set to 2). FCM uses the following optimization steps to reach the optimal situation:

$$U^{(t+1)} = argmin \, J_m^{(FCM)}\{U, V^{(t)}\} \quad (7)$$

$$V^{(t+1)} = argmin \, J_m^{(FCM)}\{U^{(t+1)}, V\} \quad (8)$$

where *t* represents the number of iteration steps. $V^{(0)}$ and $U^{(0)}$ are initiated randomly, and their values are updated through the optimization procedure. The following equations calculate the membership function values and the vector of the cluster centers:

$$u_{ij}^{(t+1)} = \left(\sum_{k=1}^{c} \left(\frac{\left\| x_j - v_i^{(t)} \right\|_2^2}{\left\| x_j - v_k^{(t)} \right\|_2^2} \right)^{\frac{1}{m-1}} \right)^1 \quad (9)$$

$$v_i^{(t+1)} = \frac{\sum_{j=1}^{n} \left(u_{ij}^{(t+1)}\right)^m x_j}{\sum_{j=1}^{n} \left(u_{ij}^{(t+1)}\right)^m}$$

C. Fuzzy sets and fuzzy logic systems

Fuzzy set theory was first introduced by Zadeh [22]. He proposed this logic to simulate uncertain situations in the human brain by using a membership function between zero and one. The significant difference between fuzzy logic and crisp concepts is the Boolean concept that a particular object may or may not have a specific value. But, in fuzzy logic, a particular value is given a range from 0 to 1. This membership function helps experts to define linguistic variables for the inputs and outputs of their systems.

Definition 1. Fuzzy set [23]: A is a fuzzy set and represents a universal set X by a membership function. $\forall \, x \in X, \mu_A(x) \in [0,1], i.e \; \mu_A(x): X \to [0,1]$.

Definition 2. A-cut [23]: The α-cut or α-level set of the fuzzy set A is the crisp set A_α defined by:

$$A_\alpha = \{x \epsilon X \, | \mu_A(x) \geq \alpha\}$$

Definition 3. Fuzzy number [23]: A fuzzy set A in $\mathbb{R}$ satisfies the following conditions:

- A is normal,
- A_α is a closed interval for every $\alpha \in (0,1]$,
- The support of A is bounded.

Definition 4. Fuzzy logic system (FLS) [24]: A simple fuzzy logic system consists of three phases: 1) fuzzification; 2) a fuzzy interface engine; and 3) defuzzification. The first step transforms crisp inputs and variables into fuzzy sets. Then, a fuzzy interface engine defines the relationship between fuzzy input and output variables and, finally, the fuzzy output variable is transformed into a crisp output through defuzzification.

D. Data-Driven Decision Support Systems

Lu et al. [1] categorized DSSs into two main categories: traditional decision support systems, and data-driven decision support systems. Within both these categories, we have multi-criteria decision-making (MCDM), which is one of the first model-driven DSS. In MCDMs, various decision-making techniques are used to find the best option for solving a problem, such as simple linear weighing, TOPSIS (the technique for order of preference by similarity to ideal solution), analytic hierarchy processes (AHPs), and so on [25, 26]. Moreover, fuzzy versions of these techniques have been used to model uncertainty in complex situations [27, 28]. For instance, Seiti et al. [29] proposed a risk-based fuzzy DSS that involves a fuzzy MCDM approach to assess the failure of components and equipment. The fuzzy approach was chosen due to a lack of available information to apply quantitative models. They used fuzzy numbers to explain reliability, associated risks, and error for analyzing metrics. They evaluated the effectiveness of the proposed method across different scenarios in a steel plant case study, and the results gave flexibility and confidence for decision-makers to handle risks in uncertain situations. In another study proposed by Zhu et al. [30], uncertainty during the decision-making process was suitably handled. They presented a fuzzy rough number for design concept evaluation and used this concept in two methodologies, including fuzzy AHP and fuzzy TOPSIS. The results showed that the fuzzy rough number had outstanding performance in group decision making. These traditional forms of decision-making require a set of options and criteria to rank alternatives according to the goal of a decision-maker.

Moreover, multiple criteria should be defined, experts and stakeholders should answer the related questions, and numerical values are processed to select or classify one choice [3].

The second category is data-driven DSS. By integrating diverse operational databases with data warehouse technology, structured data has been widely used for supporting decisions [31]. These integrated data have invaluable information about the future to make better data-driven decisions. Further, these data can be stored as both internal and external information available through transactional systems or the Internet in an integrated data warehouse, which plays a useful role in data-driven decision making [32].

E. Fuzzy Decision Support Systems

The first time that the application of fuzzy logic and fuzzy set theories were found in decision analysis and DSS was in the early 1970s. [36]. Since that time, various applications of fuzzy logic have been proposed to handle uncertain situations of decision-making processes [34, 37, 38].

Fuzzy risk assessment is widely used to apply data-driven DSS for risk evaluation. Namvar et al. [12] proposed a data-driven DSS to assess the risk of lenders in financial service companies. They proposed a machine learning framework in a peer-to-peer lending environment. Their results show that supervised learning machine learning models can help decision-makers with risk assessment in banking by automatically scoring credit risk. Another study [33], proposed an intelligent situation awareness support system to manage abnormal situations, including hardware failure and human mistakes. They assessed the risk of abnormal events using Bayesian networks and fuzzy logic in a safety-critical environment.

Recent advancements in information systems and big data on one hand, and advancements in artificial intelligence and machine learning algorithms, on the other hand, provide new opportunities for decision-makers to use big-data-driven DSSs in more innovative ways [35]. Sensor data is noisy and providing analytical results from these data might be faulty or uncertain. Therefore, we used fuzzy logic to reduce the impact of these noisy data in big data environments. Currently, most of the studies on big data-driven DSSs that have been proposed are based on the capabilities of the supervised learning algorithms and labeled data [39], but labeled data is not always easily accessible in the real-world. Moreover, according to a study by Lu et al. [1], using unsupervised learning techniques in a data-driven DSS is still cause for concern. In addition, Shukla et al. [40] advises that, although extensive research has been carried out on big data-driven DSSs, few studies exist that apply fuzzy logic to reduce the uncertainty in big data. To cover these gaps, we propose an AFDSS which leverages an unsupervised learning algorithm and fuzzy logic for risk assessment.

III. Autonomous Fuzzy Decision Support System (AFDSS)

The risks associated with unwanted situations or events are calculated based on probability and severity. Hence, risk assessment can be considered a process of estimating those two variables. We took those principles and developed a framework that identifies and assesses risk. The framework has two main components, as illustrated in Fig.1. The risk identification component extracts the risk criteria from big data and stores them in a knowledgebase. The risk assessment component calculates the risk level of new data according to the similarity between the extracted criteria and their risk level.

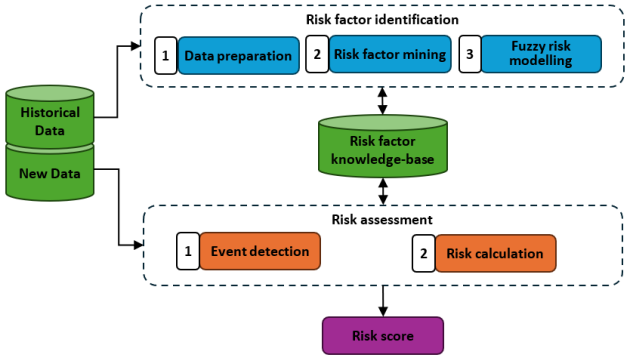

Fig 1- Autonomous Fuzzy Decision Support System (AFDSS) framework illustrating the extraction of risk criteria from historical data and subsequent intake of new data for risk assessment.

A. Risk Factor Identification

The workflow in this component starts with data preparation. The risk factors and then mined and the risk levels of those factors are then estimated.

1) Data Preparation

Data preparation is an essential step in any data mining or knowledge discovery project. Hence, the primary goal of this step is to clean the data and reduce its complexity. The data generated by IoT devices have big data characteristics, i.e., they are multi-dimensional, large-scale, and fast.

Consider a multi-dimensional data of n dimensions as $D = \{D_1, D_2, \dots, D_n\}$, where each dimension D_i represents a time series of data with similar length. Further, $x \in D_i^j$, which is equal to the j-th value of the i-th dimension. For example, we have three dimensions in our case study. $D = \{V, A_x, A_y\}$ – velocity, x-axis acceleration, and y-axis acceleration. Each dimension shows the value of instantaneous velocity, x-axis or y-axis acceleration/deceleration over time t [41].

To prepare the input data for analysis, it needs to be divided into time windows of similar length, i.e., $W = \{W_1, W_2, \dots, W_n\}$, where W_s is the s-th sequence of data with a number of values depending on the length of window. So, if we had an input data of, say, length 4096, and the length of each time window was equal to 128, then we would have 32 windows in total [41].

Our data preparation algorithm comes from previous work [42]. Designed for pattern recognition with big data of IoT - generated data, the algorithm divides the input data into fixed-length segments. Each window contains three-dimensional data of a fixed length. Then a relative unconstrained least-squares change detection method detects the events with the highest rate of change according to three key characteristics. Once finalized, the divided data form a dataset that can be used for the

subsequent steps.

2) Risk Factor Mining

Defining the criteria for decision-making is a critical step in any risk identification system. A complete list of all risk and success criteria is typically defined according to the literature and/or expert judgments. However, this process is both costly and time-consuming [43]. So, to remove this process, freeing up time and financial resources, we propose a two-stage clustering algorithm that extracts unique patterns. These unique patterns play the role of criteria in our DSS. The model categorizes various time windows into similar groups through a two-stage clustering technique that involves SOM then fuzzy k-means clustering.

- *SOM*

The SOM algorithm reduces the complexity of data to a subspace with two dimensions and transforms the input data into a map of M×N neurons. The size of the map depends on the number of input records in the source data [44]. SOM iteratively maps the input records to the closest neuron in the hidden layers of the feature map. This neuron is known as the best matching unit (BMU). Then, the weight vector of each neuron is updated according to this change. The process is repeated until no remarkable changes in the data are detected. The advantage of using the SOM algorithm for clustering is that data generated from sensors is usually large-scale and noisy. A dimension reduction method decreases both the computational cost and the impact of the noise [16].

- *Fuzzy clustering*

Real-world problems and the data that they manifest are a little different from academic research and laboratory-generated data. For example, one category of data generated by IoT devices can be similar to more than one group of data. Thus, we used fuzzy clustering to calculate one score to find the similarity between the input data and the selected clusters [45].

Although the SOM algorithm in the first stage has reduced the dataset into an abstract subspace, there are still too many points to analyze directly. Thus, these abstracted subspaces need to be further categorized into similar groups. Here, a fuzzy clustering algorithm extracts unique patterns from them. To cluster the abstract data X from the SOM, the objective function as presented in Eq. (6) is used considering that the square weighted distance is calculable by

$$\|x_k - v_i\|^2 = \sum_{j=1}^{n} \frac{(x_{kj} - v_{ij})^2}{\sigma_j^2} \tag{10}$$

where m is the fuzzification coefficient, which is greater than 1, and σ_j is the standard deviation of the j[th] feature. The input data X has n records that need to be clustered into c number of clusters. U is the partition matrix of a size $c \times n$, and V is the center of the clusters. The algorithm is given as Algorithm 1.

The fuzzy clustering results in a partition matrix U that indicates the coefficient of each record to the clusters. These results are used in the risk modelling step next.

Algorithm 1: Fuzzy clustering algorithm [45]

Input: Data set X and c number of clusters, ε very small threshold
Output: Data points with cluster label

1 ***Set the number of clusters as an input parameter.***
2 ***Initialize Uc×n as the partition matrix***
3 ***Do***
3.1 $v_{ij} = \frac{\sum_{k=1}^{N} u_{ik}^{m} x_{kj}}{\sum_{k=1}^{N} u_{ik}^{m}}$
3.2 $u_{ij} = 1 / \sum_{s=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_s\|} \right)^{2/m-1}$
While $\|U_{Iter} - U_{iter-1}\| < \varepsilon$

3) Fuzzy Risk Modeling

The unique patterns extracted from the data using fuzzy clustering serve as the criteria for our DSSs. The risk factor assessment itself is done through an FLS, which simulates human reasoning. We chose the FLS because it can handle uncertain and vague variables. The procedure takes the factor probability and severity estimations into account and then produces a risk level.

The probability of an event occurring is calculated using predictive analysis. More specifically, the likelihood of each event is estimated using statistical analysis based on the frequency of previously occurring events. Conversely, the estimated severity of each event is treated as a business problem. Here, several domain experts should be convened to provide estimates of these various. In our case study, we calculated the severity of each event using previous research conducted in the field of transportation [24, 46].

To estimate the risk factor levels, the FLS uses membership functions. Several different membership functions might be used to determine the fuzzy linguistic variables, such as triangular, trapezoidal, or Gaussian. Selecting a suitable membership function fundamentally depends on the characteristics of the variable, the available information available, and expert knowledge. We opted for parametric (trapezoidal/triangular) functions, deeming these sufficient to capture the vagueness of the variables [47]. The relations between the input and output variables are defined using a risk matrix based on the expert's knowledge. Mamdani's fuzzy interface method [48] was used to implicate each rule and aggregate the input variables into a risk output. As a last step, the defuzzification process converts the set of fuzzy output risks into one crisp variable being a risk score.

B. Risk Assessment

After finding the risk score of all extracted criteria using the clustering algorithm, the next step is assessing each probable event's risk. This occurs in two steps: event detection and risk calculation.

1) Event Detection

Event detection is a component that detects the most similar patterns in the knowledgebase. In this component, fuzzy clustering provides a U partition matrix for all new events according to their similarity to the previously known events. Let U be partition matrix from fuzzy clustering:

$$U\ partition\ matrix = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_j \end{array} \\ \begin{bmatrix} u_{11} & \cdots & u_{1j} \\ \vdots & \ddots & \vdots \\ u_{i1} & \cdots & u_{ij} \end{bmatrix} \end{array} \tag{11}$$

where D_i is a multidimensional data, C_j represents the j-th

extracted cluster number, and u_{ij} is the similarity score of i-th input data to the j-th clusters. The fuzzy clustering algorithm is responsible for providing these fuzzy scores.

2) Risk Calculation

Risk calculation is the final component in our proposed DSS. Each group of input data contains many fixed-length windows, so the total risk per event is calculated according to the following equation:

$$RE_i = \mathcal{F}_k\left(\sum_{j=1}^{c} R_j u_{ij}\right), (i = 1,2,3, \dots, N, k = 1,2, \dots, c) \quad (12)$$

where $\mathcal{F}$ is the function used to calculate top k similar events, N is the total number of events we want to calculate risks for, and c is the total number of criteria, which were extracted using the clustering algorithm. R is the risk score of the event and u_{ij} is the similarity score which is calculated by fuzzy clustering.

The total risk score entirely depends on the risks (RE_i) in all sub-categories. Thus, the equation for determining the total risk score is:

$$Total\ Risk\ Score = \mathcal{AF}(RE_i), (i = 1,2,3, \dots, N) \quad (13)$$

where $\mathcal{AF}$ is the aggregation function, which could be optimistic, pessimistic, or neutral depending on one's business strategy. For example, the total risk score given an optimistic strategy might be equal to the minimum score of all events. In a pessimistic strategy, the maximum value might be considered as the total risk. With a neutral strategy, the average value of all events would be regarded as the total risk score.

C. Evaluation

Evaluation is key to assessing the confidence of the proposed DSS. We evaluated the model in this study using a sensitivity analysis through a usage-based insurance risk assessment case study in a big data environment. We closely monitored all used parameters of the models and assessed their impact on the risk score provided by our DSS.

IV. Case Study

To evaluate the performance of our proposed system, we undertook a real case study to assess the risk of drivers based on their driving behavior. Our dataset is a large-scale collection of data from a European insurance company that contains driving characteristics of over 500,000 journeys from more than 2500 drivers. Traditionally, the computational cost to process this entire dataset would be extremely high. But, according to Dong et al. [49], each person has their own unique driving pattern, so no new useful information would be gained by analyzing more than a few trips per driver. Hence, we selected the 20 longest trips per driver to include in the analysis. Thus, the final dataset contained 50,000 journeys (20 trips × 2500 drivers). Table I provides brief details about the data used.

We implement the framework in Python 2.7 on an Intel® Xeon® 3.01 GHz CPU, 64 GB of RAM, running a Linux operating system. The software platform was Anaconda 2.7. We used implemented versions of SOM [50], scikit-fuzzy, and change detection libraries [51].

Table I: Selected dataset

Trips	Drivers	Journeys per driver	Traveling time (minute)		
			Min	Average	Max
50,000	2,500	20	23:21	26:13	30:00

A. Data Preparation

We divided the driving characteristics into 15-second fixed-length time windows with a one-second sliding step because, according to a study by Zhang et al. [52], it takes around 15 seconds to complete a single driving event. RuLSIF-based change detection scores were then applied to remove unnecessary from data. The change detection model gave us a score showing the number of changes (driving maneuvers) in each window. We assessed almost 8-million-time windows to select the most important time frames. Following Lee and Jang [53], we selected the 5% with the highest RuLSIF scores to represent the most significant changes, and further selected all windows with a change score higher than a threshold of 68.598. Therefore 394,833 windows remained, each representing one driving event with 15 seconds of data.

B. Risk Criteria Mining

As outlined in the framework, we used SOM and fuzzy clustering to extract the unique driving patterns from our dataset. In SOM, designing a map with an appropriate number of prototypes is critical because a small n means the prototype is very generic, and large n will generate a very detailed map. Following Céréghino and Park [44], we defined an optimal number of nodes as equal to $5 \times \sqrt{n}$ where n is the total number of selected events. With 394,833 events, the optimal number of nodes was therefore 2814. In addition, we selected a map size of 21×134 based on the eigenvalues and eigenvectors [16]. After reducing the complexity of data via SOM, we clustered data points with a partitive clustering algorithm. Finding the optimum number of clusters is crucial here. Hence, we followed our previous study on this data to define the optimum number of clusters [42], which was 29. Table II summarizes the information about the extracted clusters.

Each cluster represents a group of drivers, and as discussed, this output serves as our decision-making criteria. In other words, understanding the risks associated with these driving patterns provides insights into the risk attached to all drivers with a similar driving pattern.

Table II- Fuzzy clustering result

Cluster number	Frequency Percentage (%)	Cluster number	Frequency Percentage (%)	Cluster number	Frequency Percentage (%)
17	16.493%	12	3.388%	15	1.803%
29	9.082%	18	3.101%	7	1.706%
13	6.682%	16	2.872%	1	1.629%
8	5.179%	3	2.428%	5	1.496%
2	4.955%	28	2.345%	19	1.371%
11	4.594%	22	2.200%	24	1.287%
27	4.545%	14	1.947%	4	0.978%
23	4.495%	9	1.875%	20	0.960%
6	4.369%	10	1.822%	21	0.728%
26	3.863%	25	1.808%		

C. *Fuzzy Risk Modelling*

Once the driving patterns have been extracted, the next step is to assess the risk. To do this, we first assessed the probability of risk attached to each pattern cluster, then we estimated the severity.

1) The probability estimation

From the criteria mining step, we found 29 driving patterns in our data set. These patterns are unique, and each has a different likelihood of occurring. The probability of each driving pattern was therefore estimated using the following equation:

$$P(E_i) = \frac{n(E_i)}{\sum_{i=1}^{c} n(E_i)} \quad (14)$$

where $n(E_i)$ is equal to number of times that the i-th driving pattern occurred, divided by the total number of driving events in our sample space.

The results show that the scores ranged between 0.7% and 16.5%, which we normalized to a range between 0 and 1 using a min-max transformation.

2) The severity estimation

Assessing the severity of an event depends on the situation and is usually conducted via an investigation process with a number of experts in risk assessment. In our case, we developed the severity analysis using previous research conducted by domain experts in the field of transportation [24, 46].

Most notably, we followed the research of Eboli et al. [54] to find the severity of each driving pattern. These researchers explored the relationship between velocity and acceleration to distinguish dangerous driving conditions and found correlations between dangerous driving patterns, instantaneous velocity, and acceleration. Based on their findings, a driver's behavior is risky when the acceleration value is greater than the threshold defined in the following equation:

$$|\bar{a}| = g.\left[0.198.\left(\frac{V}{100}\right)^2 - 0.592.\left(\frac{V}{100}\right)^2 + 0.569\right] \quad (15)$$

where $|\bar{a}|$ is the instantaneous acceleration norm, V is the value of velocity (km/h), and g denotes gravity, which is equal to 9.18 (m/s^2). According to this equation, when the value of acceleration is more than $|\bar{a}|(m/s^2)$, the driver is engaging in risky behavior.

We used the proposed equation to find the percentage of abnormal acceleration for each driving event. The consequence analysis proposed by Eboli et al. [46] assessed the frequency and severity of abnormal events. The distributon of driving severity is skewed; thus, we used a log transformation and a min-max transformation to standardize the average number of dangerous events per second to a range between 0 and 1.

3) Event risk estimation

The FLS uses the membership functions represented in Tables III-V and shown in Fig. 2. The relation between the probability and severity variables with risk is shown in Table VI. For example, if the probability is U and the severity is M, then the risk is TA. Mamdani's fuzzy inference method is used to calculate the output risk score. Table VII describes the functions used to obtain the fuzzy outcomes from the input variables. Finally, the defuzzification process transforms the fuzzy risk set into a crisp risk score.

TABLE III- PROBABILITY LINGUISTIC VARIABLES

Fuzzy set	Linguistic term	α cut	
		Level 1	Level 0
VU	Very Unlikely	0,0.1	0.3
U	Unlikely	0.3	0.1,0.5
E	Even	0.5	0.3,0.7
L	Likely	0.7	0.5,0.9
VU	Very Unlikely	0.9,1	0.7

TABLE IV SEVERITY LINGUISTIC VARIABLES

Fuzzy set	Linguistic term	α cut	
		Level 1	Level 0
N	Negligible	0, 0.1	0.3
MI	Minor	0.3	0.1,0.5
M	Medium	0.5	0.3,0.7
MA	Major	0.7	0.5,0.9
C	Catastrophic	0.9,1	0.7

TABLE V- RISK LINGUISTIC VARIABLES

Fuzzy set	Linguistic term	α cut	
		Level 1	Level 0
A	Acceptable	0	0.3
TA	Tolerable Acceptable	0.3	0,0.6
TNA	Tolerable not acceptable	0.6	0.3,0.9
NA	Not Acceptable	0.9,1	0.6,0.9

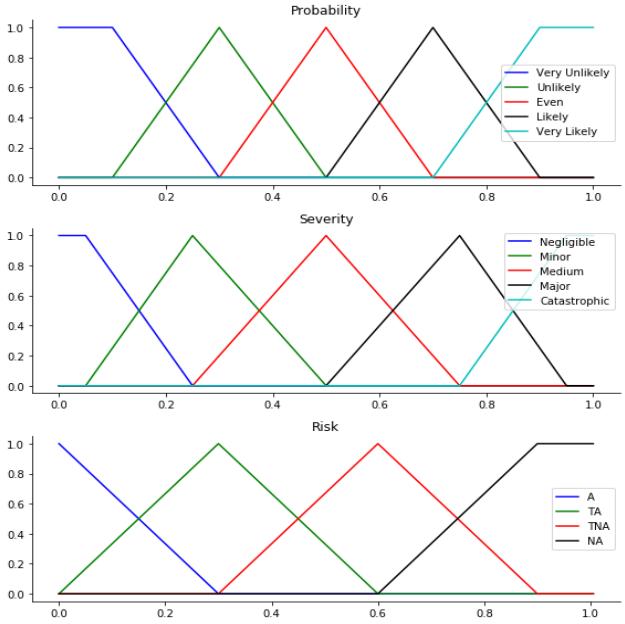

Fig 2- Fuzzy membership functions of Probability, Severity, and Risk

TABLE VI- RISK MATRIX

		Severity				
		N	MI	M	MA	C
Probability	VL	TNA	TNA	NA	NA	NA
	L	TA	TNA	TNA	NA	NA
	E	A	TA	TNA	NA	NA
	U	A	A	TA	TNA	NA
	VU	A	A	TA	TNA	TNA

TABLE VII- MAMDANI MODEL[55]

Operation	Operator	Formula
Union (OR)	MAX	$\mu_c = \max(\mu_A(x), \mu_B(x))$
Intersection (AND)	MIN	$\mu_c = \min(\mu_A(x), \mu_B(x))$
Implication	MIN	$\mu_c = \min(\mu_A(x), \mu_B(x))$
Aggregation	MAX	$\mu_c = \max(\mu_A(x), \mu_B(x))$
Defuzzification	CENTROID	$COE = Z^* = \frac{\int z\,\mu_c(z)dz}{\int \mu_c(z)dz}$

The proposed fuzzy risk estimation model has been developed to calculate the risk score of all driving events. In our previous study, we analyzed all extracted events [42]. Table VIII shows the calculated risk score for each driving category quantitatively. The results show that Clusters 20 and 21 have the top two dangerous driving events. Cluster 20 is a very high-risk cluster and accounts for 0.96% of all driving events. This cluster represents those who drive dangerously during cornering. The other high-risk cluster is Cluster 21. This cluster reflects the behavior of reckless drivers who drive faster than the speed limit. At the opposite end of the spectrum, the driving behaviors in Cluster 29 pose a low risk. The behavior clustered in this group is changing lanes at low speed.

TABLE VIII- CALCULATED RISK SCORES

Cluster Number	Probability Score	Severity Score	Risk Score
1	0.19	0.75	0.600
2	0.31	0.33	0.265
3	1.00	0	0.600
4	0.28	0.83	0.666
5	0.59	0.08	0.326
6	0.46	0.08	0.180
7	0.27	0.58	0.406
8	0.43	0.17	0.277
9	0.43	0	0.110
10	0.53	0.17	0.333
11	0.42	0.42	0.455
12	0.32	0.33	0.288
13	0.24	0.5	0.300
14	0.58	0.08	0.324
15	0.09	0.42	0.286
16	0.26	0.5	0.300
17	0.54	0.17	0.351
18	0.50	0.08	0.187
19	0.31	0.58	0.415
20	0.33	1	0.857
21	0.29	0.92	0.795
22	0.52	0.08	0.259
23	0.51	0.17	0.300
24	0.38	0.42	0.408
25	0.00	0.67	0.494
26	0.14	0	0.104
27	0.33	0.17	0.197
28	0.27	0.25	0.102
29	0.21	0.67	0.483

D. New drivers risk calculation

After calculating the risk score of all extracted driving patterns using fuzzy logic, we can calculate the risk score of new driving events according to the provided knowledge. Table IX shows the calculation procedure of one trip. In this table, the columns show the list of all extracted driving behavior in our DSS, and the rows show the list of detected driving events in one trip. The cells show the similarity score between driving events in the trip and the extracted driving patterns in the knowledgebase. The trip has 31 detected events according to the driver's behavior. We calculated the similarity score of all events with all risk factors which are extracted from our knowledge base. Afterwards, the risk score of each event was calculated according to the most similar driving behavior risk score. In this example, we calculated the risk of each event using Eq. 12. Finally. The risk score of the selected trip according to the optimistic, pessimistic, and neutral strategies are equal to 0.192, 0.646, and 0.318 respectively.

V. EVALUATION

The most important step in developing a DSS is the validation methodology. The result of this step shows the confidence of the proposed model and the provided results. The proposed DSS is based on a fuzzy clustering algorithm which extracts various criteria for decision making, and the fuzzy risk matrix is used for risk score calculation.

We developed a sensitivity analysis for various trips with different risk levels to validate the proposed system. The sensitivity analysis shows how much uncertainty in the input parameters might affect the model result. This sensitivity analysis of these drivers, which partially validates the approach, was undertaken with the following conditions:

- Condition 1: the top k similar driving patterns were selected, where all possible values for k were considered.
- Condition 2: three different aggregation strategies were defined: optimistic, pessimistic, and neutral.

We selected three trips with different risk levels and assessed their risk with the proposed DSS, considering different confidence parameters.

The driving behavior on our three trips shows that Trip A is a safe trip with no dangerous events. Trip B has few unsafe driving events, and Trip C is a very dangerous trip with numerous high-risk events. We calculated the risk score of these trips using all possible parameters in our DSS. Fig. 3 shows the sensitivity analysis results across the three different strategies. The x-axis shows the number of k according to Condition 1, and the y-axis shows the calculated risk score.

The analysis shows that the model has very stable results. Changing the models' parameters had no impact on performance. In all results, safe drivers demonstrated a lower risk score in comparison to medium and dangerous drivers using all parameters according to Conditions 1 and 2. Moreover, the figures show that variations decrease by increasing the number of selected driving patterns (k), and the model becomes more stable.

Fig. 3A shows the results for the optimistic strategy. This strategy's range of scores is very small because it optimistically ignores some of the high-risk events and bases its risk scores on the minimum calculations. Notably, the optimistic strategy is more reliable for assessing the risk of safe drivers. The pessimistic strategy, however, shows different behavior. By increasing the value of k, risk score decreases, and the risk value only becomes stable once k is reaches 9. The neutral strategy shows different behavior again. The risk score for Trip A, the safe trip, increases, but the score for the other two trips

decreases. The sensitivity analysis shows that changing the model's parameters has no impact on the final results because trip C is always dangerous, regardless of the input parameters

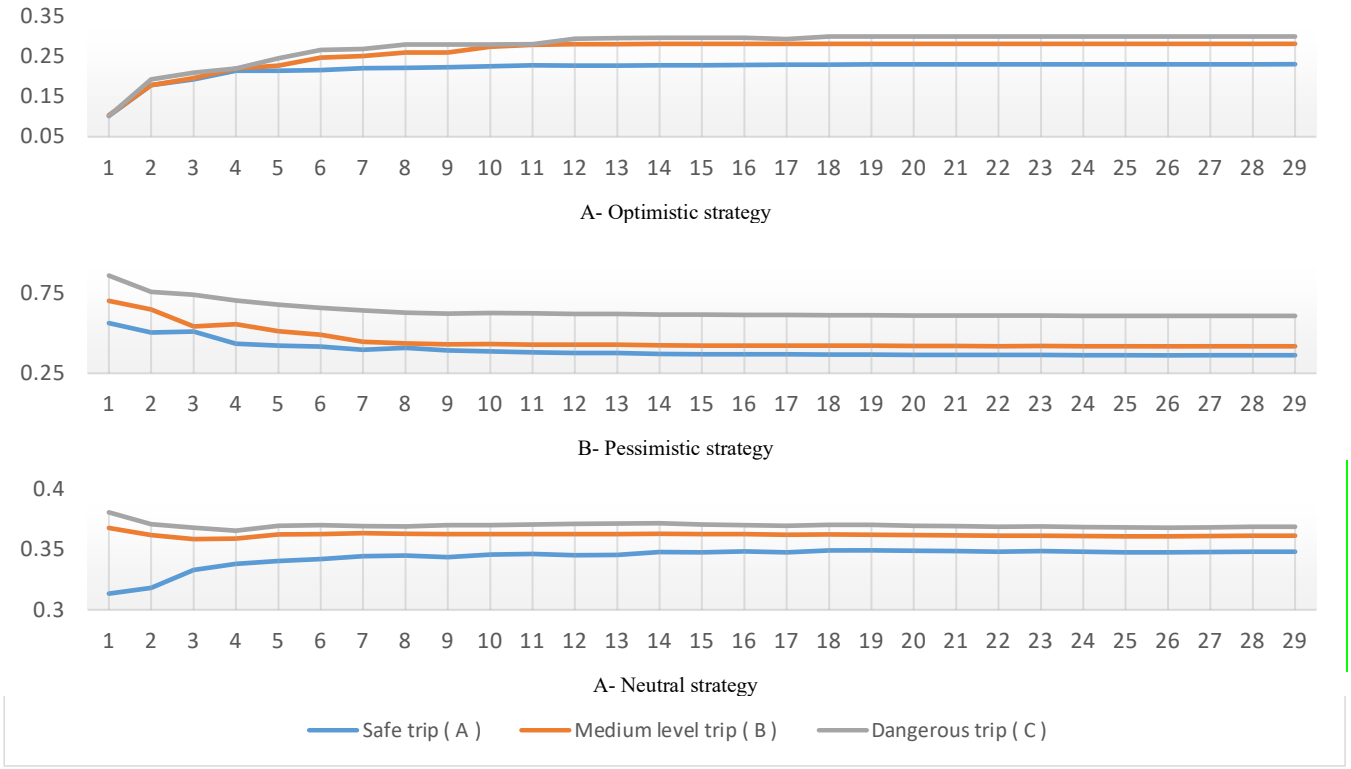

Fig 3- Sensitivity analysis strategies
Sensitivity analysis results across three different strategies are shown into three parts. Part A shows optimistic strategy and part B and C depict pessimistic and Neutral strategies.

for possible value of k in all strategies.

Selecting a suitable strategy with the optimum number of k truly depends on the type of driving behavior we want to evaluate the risk of. In this regard, a pessimistic strategy is more applicable to dangerous drivers than an optimistic one and vice versa. A neutral strategy would be most useful without any predetermined ideas of the severity of the driving risk associated with certain behaviors. The most important note in this analysis is that AFDSS allows us to compare the risk scores of two different types of drivers using the same strategy.

VI. Discussion

The proposed DSS has superior performance in comparison to traditional forms of DSS. They typically overlook the relationships among the involved criteria and are not able to identify the imprecise reasoning embedded in their criteria. Rule-based risk assessment methods are also time-consuming and challenging as they rely heavily on expert knowledge, making them very subjective. AFDSS relies upon new advancements in artificial intelligence and machine learning to provide new opportunities for risk experts to use unsupervised learning and automatic rule extraction algorithms to build risk models in various domains. The availability of labeled data is a major concern for risk modeling in a big data environment. Rule extraction with neural networks, random forest decision making models, and deep learning models all give outstanding performance in this area when labeled data is available, but, without true labels, none are not suitable for helping risk experts with analytics.

These weaknesses mean unsupervised learning techniques need to be applied to risk management to understand unknown labels in big and complex datasets. Various unsupervised learning models have been applied for risk assessment and anomaly detection. These models try to cluster unlabeled data into number groups, but they cannot provide a fuzzy membership function to cluster data. Therefore, in this study, we clustered unlabeled data into different clusters with a fuzzy membership function that shows the similarity of big data patterns to all clusters. Then, we applied fuzzy logic to assess the risk level of all extracted patterns to develop our proposed DSS.

The proposed system can contribute to the risk assessment industry in big data environments. However, using this model in a practical sense would require a particular big data platform with a cloud computing feature. Further, all the functionalities of the proposed system would need to be accessible to employees by implementing this framework on a distributed computing system.

VII. Conclusion and Future Works

This paper proposes a new DSS for risk assessment in big

data environments. To support decision-makers, we applied an autonomous artificial intelligence and machine learning algorithm to extract decision-making criteria and risk factors from big data. Our DSS automatically extracts patterns from big data, the risks associated with those patterns are assessed using a fuzzy inference system. The proposed system is evaluated via a case study on usage-based insurance risk drawn from a telematics-generated big dataset – namely, driver's insurance and driving behaviors. Our results, including a sensitivity analysis, show that the results from the proposed DSS are consistent across various input parameters and optimistic, pessimistic, and neutral business strategies.

In future, we want to implement this framework for supervised learning environments. One of the main challenges of this study was lack of labeled data, which led us to provide an unsupervised learning risk assessment method. Now that we have labeled data, we can use these labels to build a supervised model. Proposing a deep sequential model to predict the next driving behavior is another research stream that needs to be investigated. Moreover, AI and machine learning are black box models that need explainability to be useful in business. Thus, we also need to expend effort towards explainable AI in future studies to turn black box models into white boxes. Last but not least, the applications of AFDSS are not limited to insurance. Work needs to be done to extend the framework to other domains, such as cyber security, financial risk prediction, and fraud detection.

TABLE IX- TRIP RISK CALCULATION PROCESS

E\RF	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	RE
1	0.002	0.025	0.050	0.002	0.001	0.003	0.038	0.019	0.129	0.029	0.059	0.057	0.175	0.006	0.071	0.015	0.030	0.003	0.033	0.030	0.014	0.002	0.018	0.008	0.004	0.089	0.005	0.007	0.074	**0.219**
2	0.002	0.058	0.022	0.002	0.001	0.003	0.033	0.011	0.170	0.016	0.031	0.026	0.058	0.004	0.091	0.009	0.016	0.003	0.076	0.016	0.009	0.002	0.011	0.006	0.003	0.033	0.004	0.005	0.280	**0.342**
3	0.001	0.196	0.014	0.001	0.001	0.002	0.036	0.008	0.046	0.012	0.015	0.016	0.033	0.003	0.094	0.006	0.011	0.002	0.246	0.010	0.006	0.002	0.007	0.004	0.003	0.020	0.003	0.004	0.197	**0.445**
4	0.002	0.235	0.015	0.001	0.001	0.002	0.070	0.008	0.027	0.015	0.012	0.016	0.034	0.004	0.119	0.007	0.012	0.002	0.281	0.011	0.007	0.002	0.008	0.005	0.003	0.021	0.003	0.004	0.075	**0.347**
5	0.002	0.070	0.023	0.002	0.001	0.003	0.289	0.011	0.027	0.025	0.014	0.023	0.059	0.004	0.178	0.009	0.018	0.003	0.093	0.016	0.009	0.002	0.011	0.006	0.003	0.034	0.004	0.005	0.057	**0.360**
6	0.002	0.029	0.045	0.002	0.001	0.003	0.201	0.018	0.031	0.053	0.019	0.043	0.148	0.006	0.109	0.014	0.034	0.003	0.038	0.027	0.013	0.002	0.016	0.008	0.004	0.077	0.005	0.006	0.042	**0.361**
7	0.002	0.012	0.103	0.001	0.001	0.003	0.056	0.026	0.022	0.114	0.020	0.084	0.109	0.006	0.035	0.018	0.067	0.003	0.015	0.047	0.017	0.002	0.023	0.009	0.004	0.169	0.005	0.007	0.021	**0.196**
8	0.002	0.009	0.135	0.001	0.001	0.003	0.026	0.032	0.021	0.082	0.026	0.098	0.083	0.006	0.022	0.021	0.084	0.003	0.011	0.067	0.020	0.002	0.028	0.010	0.004	0.176	0.005	0.007	0.016	**0.319**
9	0.001	0.005	0.230	0.001	0.001	0.003	0.014	0.037	0.012	0.059	0.017	0.157	0.030	0.005	0.011	0.022	0.112	0.002	0.006	0.094	0.020	0.002	0.032	0.009	0.003	0.097	0.004	0.006	0.009	**0.474**
10	0.003	0.014	0.097	0.002	0.001	0.005	0.028	0.038	0.039	0.053	0.060	0.085	0.103	0.009	0.032	0.027	0.067	0.004	0.016	0.068	0.025	0.003	0.035	0.014	0.006	0.124	0.007	0.011	0.026	**0.192**
11	0.002	0.014	0.090	0.002	0.001	0.004	0.029	0.031	0.067	0.037	0.068	0.144	0.080	0.008	0.034	0.023	0.048	0.004	0.018	0.053	0.021	0.003	0.028	0.012	0.005	0.124	0.006	0.009	0.034	**0.203**
12	0.003	0.015	0.099	0.002	0.002	0.005	0.061	0.035	0.026	0.104	0.024	0.084	0.077	0.009	0.038	0.026	0.077	0.004	0.019	0.057	0.024	0.003	0.032	0.013	0.006	0.111	0.007	0.011	0.025	**0.214**
13	0.002	0.007	0.087	0.002	0.001	0.005	0.020	0.056	0.012	0.195	0.016	0.056	0.031	0.009	0.014	0.037	0.153	0.004	0.008	0.087	0.034	0.003	0.050	0.016	0.006	0.058	0.007	0.012	0.010	**0.341**
14	0.002	0.003	0.048	0.001	0.001	0.004	0.007	0.174	0.007	0.038	0.012	0.034	0.013	0.010	0.006	0.088	0.111	0.003	0.004	0.130	0.076	0.002	0.150	0.021	0.005	0.024	0.007	0.013	0.005	**0.288**
15	0.002	0.002	0.028	0.001	0.001	0.004	0.005	0.208	0.005	0.021	0.010	0.022	0.008	0.010	0.004	0.138	0.053	0.003	0.003	0.075	0.116	0.002	0.209	0.024	0.005	0.015	0.007	0.014	0.004	**0.289**
16	0.002	0.003	0.039	0.002	0.001	0.005	0.006	0.184	0.007	0.025	0.016	0.032	0.012	0.012	0.006	0.114	0.065	0.004	0.004	0.104	0.100	0.003	0.171	0.027	0.006	0.022	0.009	0.017	0.005	0.288
17	0.004	0.010	0.080	0.003	0.002	0.007	0.017	0.067	0.031	0.037	0.103	0.082	0.039	0.015	0.018	0.050	0.066	0.006	0.011	0.093	0.047	0.005	0.062	0.024	0.009	0.063	0.012	0.019	0.018	**0.646**
18	0.003	0.017	0.073	0.002	0.002	0.005	0.027	0.031	0.097	0.032	0.128	0.095	0.075	0.009	0.036	0.024	0.043	0.004	0.021	0.049	0.023	0.004	0.029	0.013	0.006	0.091	0.007	0.011	0.041	**0.306**
19	0.002	0.035	0.046	0.002	0.001	0.004	0.062	0.018	0.099	0.030	0.037	0.053	0.135	0.006	0.108	0.014	0.029	0.003	0.047	0.028	0.013	0.003	0.017	0.008	0.004	0.077	0.005	0.007	0.105	**0.294**
20	0.002	0.038	0.034	0.001	0.001	0.003	0.105	0.013	0.051	0.027	0.021	0.036	0.151	0.004	0.216	0.010	0.022	0.002	0.053	0.020	0.010	0.002	0.012	0.006	0.003	0.060	0.004	0.005	0.089	**0.291**
21	0.002	0.053	0.031	0.002	0.001	0.003	0.070	0.013	0.063	0.025	0.025	0.032	0.134	0.005	0.202	0.011	0.021	0.003	0.074	0.020	0.010	0.002	0.013	0.007	0.003	0.052	0.004	0.005	0.113	**0.291**
22	0.002	0.079	0.020	0.001	0.001	0.003	0.038	0.010	0.097	0.015	0.023	0.022	0.055	0.004	0.117	0.008	0.014	0.002	0.108	0.014	0.008	0.002	0.009	0.005	0.003	0.030	0.003	0.005	0.300	**0.427**
23	0.001	0.203	0.014	0.001	0.001	0.002	0.037	0.008	0.044	0.012	0.014	0.015	0.033	0.003	0.096	0.006	0.011	0.002	0.264	0.010	0.006	0.002	0.007	0.004	0.003	0.020	0.003	0.004	0.174	**0.350**
24	0.001	0.507	0.006	0.001	0.000	0.001	0.020	0.003	0.014	0.006	0.006	0.006	0.014	0.002	0.044	0.003	0.005	0.001	0.293	0.004	0.003	0.001	0.003	0.002	0.001	0.008	0.001	0.002	0.041	**0.320**
25	0.001	0.563	0.005	0.001	0.000	0.001	0.017	0.003	0.012	0.005	0.005	0.005	0.012	0.001	0.037	0.002	0.004	0.001	0.268	0.004	0.002	0.001	0.003	0.002	0.001	0.007	0.001	0.001	0.035	**0.314**
26	0.000	0.641	0.004	0.000	0.000	0.001	0.012	0.002	0.009	0.004	0.004	0.004	0.009	0.001	0.028	0.002	0.003	0.001	0.229	0.003	0.002	0.001	0.002	0.001	0.001	0.005	0.001	0.001	0.028	**0.305**
27	0.000	0.702	0.003	0.000	0.000	0.001	0.010	0.002	0.007	0.003	0.003	0.003	0.007	0.001	0.022	0.002	0.003	0.001	0.195	0.002	0.001	0.000	0.002	0.001	0.001	0.004	0.001	0.001	0.021	**0.298**
28	0.000	0.750	0.003	0.000	0.000	0.000	0.009	0.002	0.006	0.002	0.002	0.003	0.006	0.001	0.019	0.001	0.002	0.000	0.164	0.002	0.001	0.000	0.001	0.001	0.001	0.004	0.001	0.001	0.018	**0.292**
29	0.000	0.740	0.003	0.000	0.000	0.000	0.009	0.002	0.006	0.003	0.003	0.003	0.006	0.001	0.019	0.001	0.002	0.000	0.171	0.002	0.001	0.000	0.001	0.001	0.001	0.004	0.001	0.001	0.019	**0.293**
30	0.000	0.879	0.001	0.000	0.000	0.000	0.004	0.001	0.003	0.001	0.001	0.001	0.003	0.000	0.009	0.001	0.001	0.000	0.082	0.001	0.001	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.008	**0.278**
31	0.000	0.881	0.001	0.000	0.000	0.000	0.004	0.001	0.003	0.001	0.001	0.001	0.003	0.000	0.008	0.001	0.001	0.000	0.081	0.001	0.001	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.008	**0.278**

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council through the Laureate Fellow Project under Grant FL190100149.

REFERENCES

[1] J. Lu, Z. Yan, J. Han, and G. Zhang, "Data-Driven Decision-Making (D 3 M): Framework, Methodology, and Directions," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 3, pp. 286-296, 2019.

[2] D. J. Power and R. Sharda, "Model-driven decision support systems: Concepts and research directions," Decision Support Systems, vol. 43, pp. 1044-1061, 2007.

[3] E. Mulliner, N. Malys, and V. Maliene, "Comparative analysis of MCDM methods for the assessment of sustainable housing affordability," Omega, vol. 59, pp. 146-156, 2016.

[4] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, and A. Hussain, "Big data and IoT-based applications in smart environments: A systematic review," Computer Science Review, vol. 39, p. 100318, 2021.

[5] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information sciences, vol. 275, pp. 314-347, 2014.

[6] W. Li, Y. Chai, F. Khan, S. R. U. Jan, S. Verma, V. G. Menon, et al., "A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system," Mobile Networks and Applications, pp. 1-19, 2021.

[7] R. Audi, "The Cambridge dictionary of philosophy," 1999.

[8] A. Sengupta, D. Bandyopadhyay, C. Van Westen, and A. Van Der Veen, "An evaluation of risk assessment framework for industrial accidents in India," Journal of loss prevention in the process industries, vol. 41, pp. 295-302, 2016.

[9] D. Liu and T. J. Stewart, "Integrated object-oriented framework for MCDM and DSS modelling," Decision Support Systems, vol. 38, pp. 421-434, 2004.

[10] Y. Zhang, P. Geng, C. Sivaparthipan, and B. A. Muthu, "Big data and artificial intelligence based early risk warning system of fire hazard for smart cities," Sustainable Energy Technologies and Assessments, vol. 45, p. 100986, 2021.

[11] J. Loy-Benitez, P. Vilela, Q. Li, and C. Yoo, "Sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks," Ecotoxicology and environmental safety, vol. 169, pp. 316-324, 2019.

[12] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," International Journal of Computational Intelligence Systems, vol. 11, pp. 925-935, 2018.

[13] A. A. Taha and S. J. Malebary, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," IEEE Access, vol. 8, pp. 25579-25587, 2020.

[14] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," Journal of Big Data, vol. 6, p. 44, 2019.

[15] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, vol. 78, pp. 1464-1480, 1990.

[16] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," IEEE Transactions on neural networks, vol. 11, pp. 586-600, 2000.

[17] H.-k. Du, J.-x. Cao, Y.-j. Xue, and X.-j. Wang, "Seismic facies analysis based on self-organizing map and empirical mode decomposition," Journal of Applied Geophysics, vol. 112, pp. 52-61, 2015.

[18] P. Mangiameli, S. K. Chen, and D. West, "A comparison of SOM neural network and hierarchical clustering methods," European Journal of Operational Research, vol. 93, pp. 402-417, 1996.

[19] H. Zhang, T. W. Chow, and Q. J. Wu, "Organizing books and authors by multilayer SOM," IEEE transactions on neural networks and learning systems, vol. 27, pp. 2537-2550, 2016.

[20] J. Heil, V. Häring, B. Marschner, and B. Stumpe, "Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils," Geoderma, vol. 337, pp. 11-21, 2019.

[21] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, p. 4, 2007.

[22] L. A. Zadeh, "Fuzzy sets," Information and control, vol. 8, pp. 338-353, 1965.

[23] A. F. Shapiro, "Fuzzy random variables," Insurance: Mathematics and Economics, vol. 44, pp. 307-314, 2009.

[24] M. Siami, A. Namvar, M. Naderpour, and J. Lu, "A fuzzy telematics data-driven approach for vehicle insurance policyholder risk assesment," in Data Science and Knowledge Engineering for Sensing Decision Support. vol. Volume 11, ed: WORLD SCIENTIFIC, 2018, pp. 1407-1414.

[25] C. Bao, D. Wu, and J. Li, "A Knowledge-Based Risk Measure From the Fuzzy Multicriteria Decision-Making Perspective," IEEE Transactions on Fuzzy Systems, vol. 27, pp. 1126-1138, 2018.

[26] S. Midya, S. K. Roy, and F. Y. Vincent, "Intuitionistic fuzzy multi-stage multi-objective fixed-charge solid transportation problem in a green supply chain," International Journal of Machine Learning and Cybernetics, vol. 12, pp. 699-717, 2021.

[27] A. C. Tolga, I. B. Parlak, and O. Castillo, "Finite-interval-valued Type-2 Gaussian fuzzy numbers applied to fuzzy TODIM in a healthcare problem," Engineering Applications of Artificial Intelligence, vol. 87, p. 103352, 2020.

[28] S. K. Das, S. K. Roy, and G.-W. Weber, "Application of Type-2 Fuzzy Logic to a Multiobjective Green Solid Transportation–Location Problem With Dwell Time Under Carbon Tax, Cap, and Offset Policy: Fuzzy Versus Nonfuzzy Techniques," IEEE Transactions on Fuzzy Systems, vol. 28, pp. 2711-2725, 2020.

[29] H. Seiti, A. Hafezalkotob, S. E. Najafi, and M. Khalaj, "Developing a novel risk-based MCDM approach based on D numbers and fuzzy information axiom and its applications in preventive maintenance planning," Applied Soft Computing, vol. 82, p. 105559, 2019.

[30] G.-N. Zhu, J. Hu, and H. Ren, "A fuzzy rough number-based AHP-TOPSIS for design concept evaluation under uncertain environments," Applied Soft Computing, p. 106228, 2020.

[31] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," Decision support systems, vol. 33, pp. 111-126, 2002.

[32] J. Huber, S. Müller, M. Fleischmann, and H. Stuckenschmidt, "A data-driven newsvendor problem: From data to decision," European Journal of Operational Research, pp. 904-915, 2019.

[33] M. Naderpour, J. Lu, and G. Zhang, "An intelligent situation awareness support system for safety-critical environments," Decision Support Systems, vol. 59, pp. 325-340, 2014.

[34] S. K. Roy, S. Midya, and G.-W. Weber, "Multi-objective multi-item fixed-charge solid transportation problem under twofold uncertainty," Neural Computing and Applications, vol. 31, pp. 8593-8613, 2019.

[35] G. Maity, S. K. Roy, and J. L. Verdegay, "Analyzing multimodal transportation problem and its application to artificial intelligence," Neural Computing and Applications, vol. 32, pp. 2243-2256, 2020.

[36] H. Zimmermann, "Fuzzy Decision Support Systems.," in Computational intelligence: soft computing and fuzzy-neuro integration with applications, Z. L. A. Kaynak O., Türkşen B., Rudas I.J., Ed., ed: Springer, Berlin, Heidelberg, 1998, pp. 198-229.

[37] S. K. Roy and S. Midya, "Multi-objective fixed-charge solid transportation problem with product blending under intuitionistic fuzzy environment," Applied Intelligence, vol. 49, pp. 3524-3538, 2019.

[38] P. Cortés-Antonio, I. Batyrshin, A. Martínez-Cruz, L. A. Villa-Vargas, M. A. Ramírez-Salinas, I. Rudas, et al., "Learning rules for Sugeno ANFIS with parametric conjunction operations," Applied Soft Computing, vol. 89, p. 106095, 2020.

[39] S. H. Chan, Q. Song, S. Sarker, and R. D. Plumlee, "Decision support system (DSS) use and decision performance: DSS motivation and its antecedents," Information & Management, vol. 54, pp. 934-947, 2017.

[40] A. K. Shukla, P. K. Muhuri, and A. Abraham, "A bibliometric analysis and cutting-edge overview on fuzzy techniques in Big Data," Engineering Applications of Artificial Intelligence, vol. 92, p. 103625, 2020.

[41] Y. Pitarch, A. Laurent, and P. Poncelet, "Summarizing multidimensional data streams: A hierarchy-graph-based approach," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2010, pp. 335-342.

[42] M. Siami, M. Naderpour, and J. Lu, "A Mobile Telematics Pattern Recognition Framework for Driving Behavior Extraction," IEEE Transactions on Intelligent Transportation Systems, 2020.

[43] A. A. Ahmadabadi and G. Heravi, "Risk assessment framework of PPP-megaprojects focusing on risk interaction and project success," Transportation Research Part A: Policy and Practice, vol. 124, pp. 169-188, 2019.

[44] R. Céréghino and Y.-S. Park, "Review of the self-organizing map (SOM) approach in water resources: commentary," Environmental Modelling & Software, vol. 24, pp. 945-947, 2009.
[45] Y. Shen, W. Pedrycz, Y. Chen, X. Wang, and A. Gacek, "Hyperplane Division in Fuzzy C-Means: Clustering Big Data," IEEE Transactions on Fuzzy Systems, 2019.
[46] L. Eboli, G. Mazzulla, and G. Pungillo, "How to define the accident risk level of car drivers by combining objective and subjective measures of driving style," Transportation research part F: traffic psychology and behaviour, vol. 49, pp. 29-38, 2017.
[47] M. Naderpour, J. Lu, and G. Zhang, "A fuzzy dynamic bayesian network-based situation assessment approach," in Fuzzy Systems (FUZZ), 2013 IEEE International Conference on, 2013, pp. 1-8.
[48] A. Geramian and A. Abraham, "Customer classification: A Mamdani Fuzzy Inference System Standpoint for Modifying the Failure Mode and Effect Analysis based Three Dimensional Approach," Expert Systems with Applications, p. 115753, 2021.
[49] W. Dong, J. Li, R. Yao, C. Li, T. Yuan, and L. Wang, "Characterizing Driving Styles with Deep Learning," arXiv preprint arXiv:1607.03611, 2016.
[50] M. Saraee, S. Vahid Moosavi, and S. Rezapour, "Application of Self Organizing Map (SOM) to model a machining process," Journal of Manufacturing Technology Management, vol. 22, pp. 818-830, 2011.
[51] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," Neural Networks, vol. 43, pp. 72-83, 2013.
[52] X. Zhang, X. Zhao, and J. Rong, "A study of individual characteristics of driving behavior based on hidden markov model," Sensors & Transducers, vol. 167, p. 194, 2014.
[53] J. Lee and K. Jang, "A framework for evaluating aggressive driving behaviors based on in-vehicle driving records," Transportation Research Part F: Traffic Psychology and Behaviour, pp. 610-619, 2017.
[54] L. Eboli, G. Mazzulla, and G. Pungillo, "Combining speed and acceleration to define car users' safe or unsafe driving behaviour," Transportation research part C: emerging technologies, vol. 68, pp. 113-125, 2016.
[55] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," IEEE transactions on computers, pp. 1182-1191, 1977.

Dr. Mohammad Siami, a senior data scientist at Optus telecom, earned his PhD from the Decision Systems and e-Service Intelligent (DeSI) Research Laboratory at the Australian Artificial Intelligence Institute (AAII), Faculty of Engineering and IT, University of Technology Sydney (UTS), Australia. With a decade-long track record, he specializes in leveraging artificial intelligence and machine learning to tackle risk assessment challenges within financial service entities, spanning banking and insurance sectors. His expertise extends to artificial intelligence, machine learning, and smartphone data analytics, reflecting his commitment to cutting-edge research in these domains.

Dr Mohsen Naderpour is an Industry Advisor and an Industry Fellow at the Australian Artificial Intelligence Institute (AAII) at the University of Technology Sydney (UTS). Mohsen began his career as a safety professional in high-risk industries, including transportation and oil, before taking up a position in academia as a research fellow with the Global Big Data Technologies Centre at UTS and then Lecture and Senior Lecturer at the Faculty of Engineering and IT. His research areas include applied artificial intelligence, risk engineering, computational intelligence, and decision support systems.

Dr. Fahimeh Ramezani, a Senior Lecturer at the School of Computer Science at the University of Technology Sydney (UTS), also serves as a core member of the Artificial Intelligence Institute (AAII). With seven years of prior industry experience as a software developer and researcher across diverse domains, her expertise enriches her academic pursuits. Her research focuses on cloud/fog computing, optimization, and artificial intelligence, reflecting her dedication to advancing these areas within academia and industry alike.

Distinguished Professor Jie Lu (F'18) AO (Officer of the Order of Australia), renowned for her groundbreaking work in computational intelligence, has made significant contributions in fuzzy transfer learning, concept drift, decision support systems, and recommender systems. With a distinguished career marked by prestigious fellowships and awards, including IEEE Fellow, IFSA Fellow, and Australian Laureate Fellow, she has published extensively and secured substantial research funding. Currently serving as Director of the Australian Artificial Intelligence Institute (AAII) at UTS and Associate Dean (Research Excellence), she leads a team of 35 researchers and 230 PhD students, driving forward cutting-edge research in the field. Her impact extends beyond academia, shaping the way organizations utilize data for decision-making in complex scenarios. With a PhD from Curtin University and a legacy of nurturing research excellence, Professor Lu's influence reverberates both nationally and internationally.