Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Using Intuitionistic Fuzzy Set to Classify Uncertain and Linearly Non-Separable Data

Shubair A. Abdullah
Department of Instructional & Learning Technology, Sultan Qaboos University, Muscat, Sultanate of Oman
Corresponding Author: Shubair A. Abdullah, **E-mail**: shubair@squ.edu.om

| ABSTRACT

The problem of non-linearly separable data points requires more efforts to classify the data sample with high accuracy. This paper proposes a new classification approach that employs intuitionistic fuzzy sets to accurately classify non-separable datasets and to efficiently deal with uncertain labelled datasets. The dataset used contains 124 students with 9 features and 1 class for each student. First, the dataset is normalized to train and test the proposed approach. Second, the intuitionistic fuzzy sets were constructed using three features and the fuzzy model was created by calculating the equation of the straight line passing through the intuitionistic fuzzy sets of dataset classes. Finally, the classification is performed by calculating the distance between each class and the unseen sample that is subject to classification. Experimental results show that the classification performance of the proposed approach is competitive and superior to that of other state-of-the-art algorithms on the aforementioned dataset.

| KEYWORDS

Data Mining, Machine learning, Fuzzy Login, Intuitionistic Fuzzy Set, Distance Measurement

| ARTICLE INFORMATION

ACCEPTED: 22 May 2024 **PUBLISHED:** 08 June 2024 **DOI:** 10.32996/jcsts.2024.6.2.12

1. Introduction

The aim of data mining algorithmes is to discover valuable information from a given data. For example, the classification of students into categories based on their motivation during e-learning classes, prediction of electric load, and detection of credit card fraud (Ricciardi et al., 2020; Navas de Maya et al., 2022). In data mining, the dataset contains a number of instances. Each instance comprises the values of a number of attributes. There are two types of datasets, labelled dataset and unlabeled datasets. In the labelled datasets, specifically designated attributes are used aiming at predicting or classifying the attribute values of instances that were unseen before. The data mining algorithm that uses labelled dataset is known as supervised learning algorithm. On the other hand, the unlabeled dataset uses attributes not specifically designated and the data mining algorithm in this case is known as unsupervised learning algorithm. The overall goal of unsupervised learning algorithms is to extract some required information from the available dataset (Alloghani et al., 2020).

For classification task, the dataset instances are divided into two parts: a training dataset and a test dataset. Various dividing strategies for the dataset have been implemented in the literature but the most general strategy is to use 70% as a training dataset and 30% as a test dataset (Xu and Goodacre, 2018). The training dataset constitutes a model of data mining algorithm that will be used to classify or predict unseen instances. A training dataset can be represented by a table in which each row is allocated to a single instance of the dataset showing the values of a number of instance attributes and the corresponding classification. The test dataset can be represented using a table similar to the one used to represent the training dataset with one difference being that the instance corresponding classification is not included. Hiding the classification in the test dataset representation in order to use it for validating the trained data mining algorithm model through predicting the ungiven classification.

Several research studies designed to investigate the latest data mining techniques have confirmed that data mining is a rapidly growing field and has been successfully applied in many fields such as Medicine (Islam et al., 2018), the Construction Industry (Yan et al., 2020), the Education (Bakhshinategh et al., 2018), the Finance and Business (Kunnathuvalappil Hariharan, 2018), and the Computer Network Security (Ahmad, Jian and Anwar Ali, 2018). However, there are still some issues that need further investigation and research. These primarily include low sample size, noisy or heterogeneous samples, class imbalance issues, computation power requirements, uncertainties of real-world data, and classification of linearly non-separable dataset. Although all these issues do require researchers' attention, the issues of uncertainty in real-world datasets and classification of linearly non-separable dataset are the prime concern of this study and represent the research problem. These two issues are common in many datasets and result from the vagueness that accompanies the description of a particular problem of a qualitative nature. On the other hand, the linearly non-separable data is data that if graphed in two dimensions, it cannot be separated by a hyper plane. Some obvious examples of uncertain dataset situations may include dataset of internet traffic and dataset of students' motivation and engagement level (Zhang, 2014; AYDİLEK, 2018).

The Intuitionistic Fuzzy Set (IFS) (Atanassov, 1986) is a very powerful tool for processing vague information. An IFS is a fuzzy set whose elements have a representation of three values: degree of membership, degree of non-membership, and the hesitation factor. By using the IFS, experts will be able to describe situations with uncertain datasets using linguistics terms (Zhang, 2014). The main goal of this study is to employ fuzzy logic to resolve the issues of uncertainty of linearly non-separable in real-world datasets. It introduces a new approach, named as Data Mining - IFS classification (DM-IFSC) approach that employs IFS to build a data mining algorithm able to accurately classify non-separable dataset and to efficiently deal with uncertain labelled datasets. The DM-IFSC is validated using real-world educational non-separable dataset with uncertainty features. Moreover, the validation covered multiclass and binary classification problems. The results show that the classification and prediction performance of the proposed model is very promising and able to deal with the problem of uncertainty caused by the qualitative nature of datasets. The rest of the paper is organized as follows. Section 2 presents background information and reviews some related works. Section 3 explains the DM-IFSC approach. Section 4 demonstrates the experiments results and Section 5 discusses them. Lastly, Section 6 concludes the paper and suggests the future directions.

2. Literature Review

As a result of the rapid technological development in the means of communication along with the advances in storage technology, numerous data are being generated and saved. These data, if properly managed, considerable amount of valuable knowledge can be extracted through the use of data mining algorithm, which are seen as reliable tools in this regard. Therefore, the use of data mining processes in supporting the decision-making has grown considerably in recent years. Some sectors such as IT related sector, educational sector, health sector, and financial sector, have adopted data mining as a general and enterprise-wide practice and have witnessed a significant increase in the application of data mining (Raja and Pandian, 2020; Jimenez-Carvelo and Cuadros-Rodríguez, 2021; Dai, Wang and Chang, 2022).

Data mining algorithms are part of Artificial Intelligence (AI) and can be classified into three categories: Machine Learning (ML), Neural Networks (NNs), and Deep Learning (DL). The ML is a subcategory of AI that enables the computer systems to learn from data observations. The Support Vector Machines (SVM), decision trees, Bayes learning, and k-means clustering are public examples of ML techniques. The Neural Networks (NNs) are a subcategory of ML that can be described as a set of connected units (neurons) usually organized in at least three layers, input, hidden, and output layer. The Deep Learning (DL) is a subcategory of NNs that includes more than one hidden layer and forms the computational multi-layer NN (Nguyen et al., 2019).

The fuzzy logic has received a lot of focus by the researchers since its inception by Lotfi Zadeh (1965), and one of the most addressed topics by the research in this field is how to integrate the fuzzy logic into the work of machine learning to resolve the issues of uncertainty of linearly non-separable in real-world datasets (Peker, 2016). Table 1 briefly shows some of the most recent related works that were published during the last few years, which are reviewed in more detail in the following.

The Fine-Tuning Fuzzy kNN (TFKNN) classifier is based on uncertainty membership (Salem et al., 2022). The classifier uses fuzzy k-Nearest Neighbors (kNN) method and modifies the membership functions based on the uncertainty theory. In order to address the hyperparameters that reduce the accuracy of the classifier, a grid search method is applied to fine-tune the fuzzy kNN method. The work introduced by Murthy et al. (2021) included employing the fuzzy logic in the image segmentation to enhance the brain tumor diagnosis process. Their classifier, named as Adaptive Fuzzy Deformable Fusion (AFDF)-based Segmentation, is created by merging fuzzy C-Means Clustering (FCM) and snake deformable approach. The classifier proposed by Ren et al. (2022) combined fuzzy theory with two machine learning algorithms, decision tree and K-means++. The combination results in a hybrid technique to overcome the ambiguity and uncertainty of logging parameters in lithology identification. Triangular fuzzy membership function is used to fuzz the logging data according to clustering center points obtained by applying the K-means++ clustering algorithm on logging data. The classification is done through a fuzzy decision tree lithology identification model. The problem of impractically transferring the data from input space to feature space in order to classify them by using the SVM algorithm has been addressed by Shojae Chaeikar et al. (2020) through the use of a Gaussian data distribution kernel, three-sigma rule, and a polygon fuzzy membership function. A comparison of the system, which is called Polygonal fuzzy weighted (PFW), radial basis function (RBF) and conventional linear kernels in identical experimental conditions showed that it could produce a high rate classification accuracy than these two commonly used kernels with SVM. A fuzzy oblique decision tree (FODT) algorithm is proposed by Cai et al. (2019). The proposed algorithm was based on axiomatic fuzzy set (AFS) in which the fuzzy rules are used to construct leaf nodes for each class in each layer of the tree. A sample that cannot be covered by the fuzzy rules are then put into an additional node, which is the only node non-leaf node.

Table 1 Some related works published between 2015-2022

Model	Integration	Year
Fine-Tuning Fuzzy kNN (TFKNN) (Salem *et al.*, 2022)	Fuzzy membership and k-Nearest Neighbors (kNN) machine learning	2022
Adaptive Fuzzy Deformable Fusion (AFDF)-based Segmentation (Murthy, Koteswararao and Babu, 2021)	C-Means Clustering (FCM) and snake deformable approach	2021
(Ren *et al.*, 2022)	Triangular fuzzy membership, decision tree and K-means++	2022
Polygonal fuzzy weighted (PFW) (Shojae Chaeikar *et al.*, 2020)	Polygon fuzzy membership function and Support Vector Machine (SVM)	2020
Oblique Decision Tree (FODT) Algorithm (Cai *et al.*, 2019)	Axiomatic fuzzy set and Decision Trees	2019
(Deborah *et al.*, 2015)	Gaussian Membership Function and Felder Silverman Learning	2015
kNN-Based Dynamic Evolving Fuzzy Clustering Method (kEFCM) (Abdulla and Al-Nassiri, 2015)	Fuzzy membership and k-Nearest Neighbors (kNN) machine learning	2015
(Subhashini *et al.*, 2022)	fuzzy logic and Convolutional Neural Network (CNN)	2022
(Thakare *et al.*, 2022)	Conventional Fuzzy Functions and Deep Learning Artificial Neural Network (ANN)	2022
CovNNet (Ieracitano *et al.*, 2022)	Fuzzy Membership Function and Deep Learning Convolutional Neural Network (CNN)	2022

When it was tested, the experimental results demonstrated an outperformance of FODT over other decision trees in terms of classification accuracy and tree size. A fuzzy rule-based system has been proposed by Deborah et al. (2015) to handle the uncertainty in the data of students of C programming language course that has been collected from their profile information and activities in the e-learning system. The system used the Gaussian membership function based fuzzy logic, and was applied to predict the learning style of 120 students and showed a significant improvement of prediction accuracy. The kNN-Based Dynamic Evolving Fuzzy Clustering Method (kEFCM) introduced by Abdulla & Al-Nassiri (2015) as a preprocessor for the neural fuzzy inference mode (Shubair, Ramadass and Altyeb, 2014). kEFCM is similar to the methods presented, which combined kNN algorithm with the fuzzy logic with the addition of an advantage of evolving the knowledgebase on which the classification of previously unseen samples is based. The knowledgebase evolving is performed through adding the newly classified samples to the training dataset samples.

Under the paradigm of artificial neural network (ANN), several approaches have been proposed in the recent years. To deal with uncertainties in opinions of online customer reviews, Subhashini et al. (2022) proposed a three-way decision making system that integrates fuzzy logic with a convolutional neural network (CNN). Within it stages, the system firstly the positive, negative, and boundary regions are classified using fuzzy concepts. Then after, to further classify fuzzy concepts originally allocated to the boundary region, a CNN used. Another example is the work presented by Thakare et al. (2022). They introduced a system to that trains a deep multiple instance learning classifier for classification of videos based on visual information from normal and abnormal videos. The use of fuzzy logic is represented through a fuzzy aggregation process adopted to fuse the anomaly scores using a few conventional fuzzy functions. The last example to mention is the CovNNet model developed by Ieracitano et al. (2022). The CovNNet is a fuzzy logic based deep learning (DL) approach resulting from integrating the fuzzy logic with CNN to classify CXR images of patients into Covid-19 pneumonia and interstitial pneumonias. The fuzzy logic is employed to handle vagueness, ambiguities and uncertainties of CXR images. The system is trained on set of fuzzy features extracted from the images using a fuzzy membership. These fuzzy features will be presented to a CNN as additional input.

The Intuitionistic Fuzzy Set (IFS) (Atanassov, 1986) is among the several branches of fuzzy set theory that have had numerous applications over the past decade (Wu, Song and Wang, 2021). The IFS is based on the condition that the degree of non-membership and the degree of membership add up to 1. In addition, the hesitation degree is the difference between 1 and the sum of the degree of non-membership and the degree, and the existence of hesitation degree enables the IFS to depict uncertain information. Because of their advantages in modeling the uncertain information of different systems, IFSs have received great interest from researchers. It has been successfully applied in many fields, including computer and network security (Xie et al., 2021), classification of students based on their performance (Meena and Thomas, 2018), and decision making (Zhang, 2014).

3. Using Intuitionistic Fuzzy Set to Classify Linearly Non-Separable Data

This section introduces a fuzzy-based DM approach, named as Data Mining - IFS classification (DM-IFSC) approach that employs IFS to build a data mining algorithm capable of solving the uncertainty and linearly non-separable dataset. The new approach has four parts: preparation of dataset, construction of IFS, creating of the fuzzy model, and verification of classifier accuracy. The framework of DM-IFSC is shown in Figure 1.

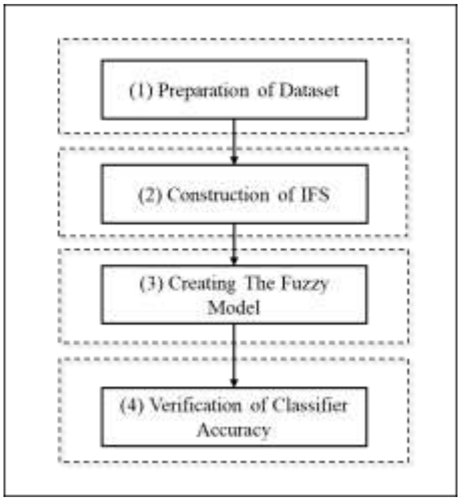
(1) Preparation of Dataset
(2) Construction of IFS
(3) Creating The Fuzzy Model
(4) Verification of Classifier Accuracy

Figure 1. Flowchart of the DM-IFSC approach

Before introducing the DM-IFSC approach, we recall the basic concepts of IFSs that have been applied in this paper.

Definition 1 (Xie *et al.*, 2021):

Let *A* an IFS in the universe of discourse $U = \{x_1, x_2, \dots \dots, x_n\}$ is defined as:

$A = \{\langle x, \mu_A(x), \nu_A(x), \pi_A(x) \rangle \ / \ x \epsilon U \}$

Where the function μ_A : U → [0,1] defines the degree of membership of $x \epsilon U$ and the function ν_A : U → [0,1] defines the degree of non-membership of $x \epsilon U$ to the set *A*. For every $x \in U, 0 \le \mu_A(x) + \nu_A(x) \le 1$. The function $\pi_A(x) = 1 - \left(\mu_A(x) + \nu_A(x)\right)$, quantifies the degree of hesitancy of $x \epsilon U$ to the set *A*.

Definition 2 (Tugrul, Gezercan and Citil, 2017)**:**

Let *A* and *B* are IFS in the universe of discourse $U = \{x_1, x_2, \dots \dots, x_n\}$ defined as:

$A = \{\langle x, \mu_A(x), \nu_A(x), \pi_A(x) \rangle \ / \ x \epsilon U \}, A \in U$

$B = \{\langle x, \mu_B(x), \nu_B(x), \pi_B(x) \rangle \ / \ x \epsilon U \}, B \in U$

A mapping of $d : U \times U$ → [0,1] will be the distance measure between *A* and *B* $d(A, B)$, if $d(A, B)$ satisfies the followings:

1) $0 \le d(A, B) \le 1$
2) $d(A, B) = 0$ if and only if $A = B$
3) $d(A, B) = d(B, A)$

3.1 Dataset Preparation

Dataset preparation is the first stage in the DM-IFSC framework. It aims to transform the raw data into acceptable input data. To achieve this, the raw data is collected from one or various sources, and it is subsequently cleaned and validated before the dataset is produced. The cleaning and validation step generally includes data digitalization and data normalization, and can include other tasks as well. Within the data digitization, the features with non-numeric values are located and converted to corresponding indices. For example, in a dataset of network traffic classification, the values of source-IP and destination-IP columns are changed to 0, 1, 2 indices. One possible outcome of this process is that we might end up with a large range of indices that cause a problem in the distribution of data. In this case, we need to normalize the values of these features to improve the convergence of values and classification accuracy.

3.2 Construction of Intuitionistic Fuzzy Sets

The second stage constructs the IFS for all classes' features from the dataset. The IFS constructed should be suitable and based on the purpose of the algorithm. Different methods for constructing the IFS have been proposed. Each method is applied in a specific case, with special requirements that may not be available in other cases. For example, the expert assessment process involved in some common methods requires an attribute metrics table and voting resolution to generate a standard metrics table (Chaira, 2019). Thus, there is a need for a general method for IFS construction that might be fit for a variety types of classification tasks and fully reflects the degrees of membership and non-membership of elements for every feature.

Before explaining the IFS construction process followed in this study, an explanation of the dataset used is given. The database is taken from records of 124 students who completed their bachelor degree in the Department of Educational Technology, College of Education, Sultan Qaboos University (SQU). Each student is represented using 9 features and 1 class. The values of the 9 features are numerical values and refer to the student's scores in eight of the high school subjects he obtained in the 12th grade, which are Arabic language, English language, Islamic studies, mathematics, physics, chemistry, biology, and sociology. The values for these scores range from 0 to 100. The seventh feature is the current cumulative grade point average (CGA), and its value is from 0 to 4. The class of student represents his graduation grade. According to the SQU's regulations, a student can get one of five grades upon graduation based on his CGA. Two tasks carried out for the purpose of preparing the training and test datasets, data cleaning and validation which involved removing rows with the blank values, and using the SQL Pivot query in MS Access to reorganize the dataset in a simple and meaningful layout that contains the fields required to train and test the DM-IFSC.

For the purpose of achieving this study objective, the degree of membership $\mu_A(x)$ and non-membership $\nu_A(x)$ were calculated on the basis that they indicate the element's tendency to be in the class and the element's non tendency to be in the class. For example, the outstanding students in Physics perform at 100, which equals 1 in the IFS representation. If student's performance is at a level of 0.75, then his performance tends to belong to the class of outstanding students with a degree of 0.75. Since there are 8 marks for each student, the average has been calculated to find the degree of membership. The value of CGA is taken as hesitation factor of IFS. This task is carried out because the mathematical model that will be created in the proposed algorithm in this research depends on the linear equation. The IFSs is calculated as follows:

$$\mu_A(x) = 1 - \left(\frac{total\ marks}{n}\right) \dots\dots\ (1)$$

$$\pi_A(x) = \ student\ CGA \dots\dots (2)$$

$$\nu_A(x) = 1 - (\mu_A(x) + \pi_A(x)) \dots\dots (3)$$

Table 2 shows five examples for five students and Table 3 shows the IFSs that has been created for the five students.

Table 2. Examples of grades for five students from the database

ARAB	ENG	ISLAM	MATH	PHYS	CHEM	BIO	SOCIO	CGA	Graduation grade
91	69	99	73	79	83	89	95	3.37	Distinction
93	82	100	67	73	79	80	100	2.91	Accept
91	74	99	91	70	81	76	96	2.43	Accept
88	94	99	77	81	80	89	94	3.34	Distinction
87	84	98	72	79	71	84	99	3.03	Accept

Table 3 Examples of IFS created for five students

$\mu_A(x)$	$\nu_A(x)$	$\pi_A(x)$	Class
0.8475	-0.1845	0.337	Distinction
0.8425	-0.1335	0.291	Accept
0.8475	-0.0905	0.243	Accept
0.8775	-0.2115	0.334	Distinction
0.8425	-0.1455	0.303	Accept

3.3 Creating the Fuzzy Model

This stage aims at modeling the features of classes by applying a data mining algorithm on the dataset prepared in the previous stage. Figure 2 shows distribution of the dataset (124 samples) and how the dataset points overlap.

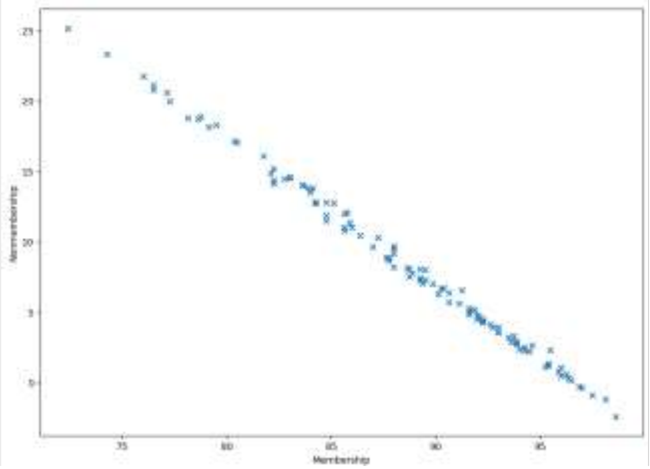

Figure 2. Distribution of the dataset (124 samples)

The process begins by fetching a class from the dataset and a set of features describe or belong to the class. Each of these features is represented by three values: membership value, non-membership value, and value of hesitancy. This produces a number of subsets of IFSs. For each class there will be a separate subset of IFSs. Then the equation of the straight line (slope) passing through the class IFSs will be calculated by using the following equations.
The linear equation in its slop-intercept form:

$$y = mx + b \ldots\ldots (4)$$

The slop (m) could be calculated by:

$$m = \left(\frac{y2 - y1}{x2 - x1}\right) \ldots\ldots (5)$$

Where $x1, y1$ and $x2, y2$ are two points in a class subset of IFSs; $x = \mu_A(x)$; $y = \pi_A(x)$.
From (4) and (5), the intercept (b) is:

$$b = y - mx \ldots\ldots (6)$$

The results from the calculations above will lead to the fuzzy model that will be used to perform the classification of unseen samples. Table 4 shows an example of fuzzy model created for the dataset used in this study. The model has five classes "Accept, Distinction, Distinction with Honors, Good, and Very Good". The linear equations of classes have been calculated and represented in the table by showing Slop (m), Intercept (b), x, and y values.

Table 4. Fuzzy model

x1	y1	x2	y2	Slope (m)	Intercept (B)	Class Name
95.5	2.35	76	21.73	-0.99385	97.26231	Accept
97.5	-0.84	82.25	14.38	-0.99803	96.4682	Distinction
98.625	-2.375	88	8.24	-0.99906	96.15718	Distinction with Honors
94.625	2.655	74.25	23.33	-1.01472	98.67325	Good
98.125	-1.155	77.25	19.98	-1.01246	98.19216	Very Good

Figure 3 summarizes the steps followed in this research to create the fuzzy model.

3.4 Classification

This stage aims at classifying unseen samples. The essence of the classification task in DM-IFSC is the calculation of the distance between each class and the unseen sample that is subject to classifying. For a given unseen sample represented by IFS, DM-IFSC algorithm calculates the distance from the unseen sample point to the classes in the dataset that are represented by a linear equation in the fuzzy model. The smallest obtained distance value gives an estimate of the nearness of the class to which the unseen sample might be classified. DM-IFSC uses the following equation to find the distance between unseen sample point and a straight line of a class:

$$d = \frac{|Ax1 + By1 + C|}{\sqrt{A^2 + B^2}} \ldots\ldots (7)$$

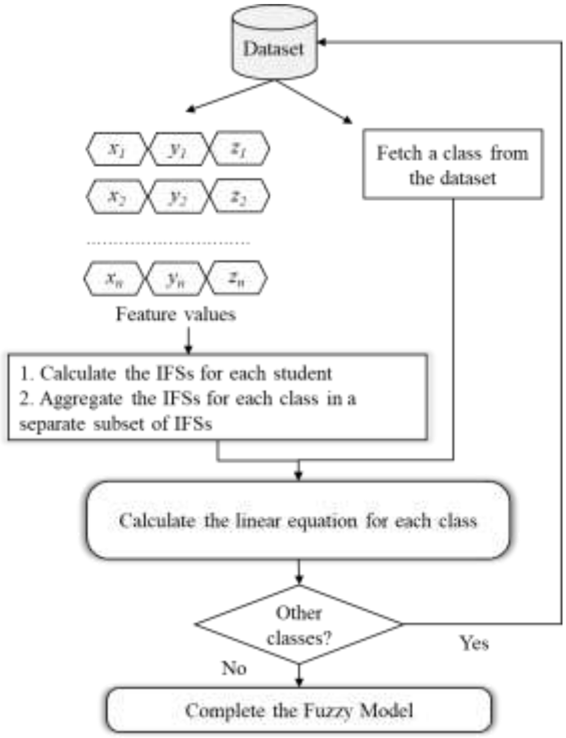
Dataset
x_1 y_1 z_1
x_2 y_2 z_2
x_n y_n z_n
Feature values
Fetch a class from the dataset
1. Calculate the IFSs for each student
2. Aggregate the IFSs for each class in a separate subset of IFSs
Calculate the linear equation for each class
Other classes?
Yes
No
Complete the Fuzzy Model

Figure 3. Steps of fuzzy model creation task

4. Experiment Results

The experimental analysis is implemented using Python on a PC with Windows 10. The hardware used was Intel(R) i7-8700 CPU 3.20GHz and a RAM of 16.0GB. The evaluation aimed to check the accuracy of the DM-IFSC in predicting students' graduation class based on their high school grades. During the evaluation, the DM-IFSC has been subjected to two types of classification problems, multiclass and binary classification problems to examine its efficiency in different situations.

4.1 Multiclass Classification

To evaluate the performance of DM-IFSC, the correct classification rate (referred to as accuracy), and incorrect classification rates (referred to as error) are calculated as follows:

$$accuracy = \frac{a}{n}$$

$$error = \frac{e}{n}$$

Where a is the number of samples that are classified correctly, e is the number of samples classified incorrectly, and n is the total of tested samples.

The proposed method was evaluated using k-fold cross-validation method (k=10) on dataset of 124 students who completed their bachelor degree in the Department of Educational Technology, College of Education, Sultan Qaboos University (SQU). The class of student represents his graduation grade. According to the SQU's regulations, a student can get one of five grades upon graduation based on his CGA. For example, if the CGA is between 3.75 - 4.00, the student will get “Distinction with Honors” grade, and if CGA is between 3.30 - 3.74, he will get “Distinction” grade. Table 5 shows the graduation average and the classes. Tables 6 shows the 10-fold cross-validation evaluation results. The results show that the overall accuracy of the DM-IFSC is 72%, which is an encouraging result if we consider the high overlapped situation of the dataset. To examine the efficiency of the proposed method, its performance on the same dataset has been compared with three machine learning algorithms that are widely used in the literature of the research field, SVM, kNN, and Naive Bayes. Tables 7-9 show the 10-fold cross-validation evaluation results the three machine learning algorithms.

Table 5. The graduation averages and the classes according to SQU regulations

Graduation average	Class
3.75 - 4.00	Distinction with Honors
3.30 - 3.74	Distinction
2.75 - 3.29	Very Good
2.30 - 2.74	Good
2.00 - 2.29	Accept

Table 6. Multiclass 10-fold cross-validation evaluation results

Test #	N	Correct	Error	Correction Rate	Error Rate
1	13	10	3	77%	23%
2	13	9	4	69%	31%
3	13	10	3	77%	23%
4	13	9	4	69%	31%
5	12	10	2	83%	17%
6	12	9	3	75%	25%
7	12	8	4	67%	33%
8	12	8	4	67%	33%
9	12	8	4	67%	33%
10	12	8	4	67%	33%
Overall				72%	28%

Table 7. Multiclass 10-fold cross-validation valuation results - SVM

Test #	N	Correct	Error	Correction Rate	Error Rate
1	13	46%	54%	46%	54%
2	13	69%	31%	69%	31%
3	13	62%	38%	62%	38%
4	13	62%	38%	62%	38%
5	12	92%	8%	92%	8%
6	12	75%	25%	75%	25%
7	12	75%	25%	75%	25%
8	12	58%	42%	58%	42%
9	12	75%	25%	75%	25%
10	12	92%	8%	92%	8%
Overall				71%	29%

Table 8. Multiclass 10-fold cross-validation valuation results - kNN

Test #	N	Correct	Error	Correction Rate	Error Rate
1	13	6	7	46%	54%
2	13	5	8	38%	62%
3	13	5	8	38%	62%
4	13	6	7	46%	54%
5	12	3	9	25%	75%
6	12	4	8	33%	67%
7	12	4	8	33%	67%
8	12	6	6	50%	50%
9	12	2	10	17%	83%
10	12	8	4	67%	33%
Overall				39%	61%

Table 9. Multiclass 10-fold cross-validation valuation results – Naive Bayes

Test #	N	Correct	Error	Correction Rate	Error Rate
1	13	6	7	46%	54%
2	13	6	7	46%	54%
3	13	7	6	54%	46%
4	13	6	7	46%	54%
5	12	4	8	33%	67%
6	12	8	4	67%	33%
7	12	5	7	42%	58%
8	12	5	7	42%	58%
9	12	4	8	33%	67%
10	12	6	6	50%	50%
			Overall	46%	54%

4.2 Binary Classification

The binary classification experiment aimed to measure the efficiency of the DM-IFSC to solve the binary classification problem. The same dataset used in the previous experiment (Multiclass Classification) used in this experiment. However, the dataset is divided into two classes only. If the average is greater of equal to 3.3 and less than or equal to 4 then the class is "Distinction", otherwise the class is "Accept". To evaluate the DM-IFSC performance, the accuracy the false alarm rates (FAR), precision, and recall were determined, the following formulas were used:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$FAR = \frac{FP}{FP + TN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{FP}{FP + FN}$$

Table 10 shows the evaluation matrix we use for TP, FP, FN, and TN.

Table 10. Evaluation matrix

		Predicted	
		Accept	Distinction
Actual	Accept	TP	FN
	Distinction	FP	TN

The performance is compared with SVM, kNN, and Naive Bayes machine learning algorithms, and the dataset is randomly divided into training and test datasets in the ratio of 7:3, i.e. 87 samples in the training dataset and 38 samples in the test dataset. Table 11 shows accuracy, FAR, precision, and recall for the DM-IFSC, SVM, kNN, and Naive Bayes.

Table 11. Accuracy, FAR, precision, and recall metrics

Algorithm	TP	FN	FP	TN	Accuracy	FAR	Precision	Recall
DM-IFSC	18	3	1	16	89%	6%	95%	86%
SVM	21	1	2	14	92%	13%	91%	95%
kNN	14	7	13	4	47%	76%	52%	67%
Naive Bayes	21	0	17	0	55%	100%	55%	100%

5. Discussion

In Tables 6-8, we can see that the accuracy of the proposed algorithm is 0.1% greater than the best accuracy obtained from the three algorithms which is SVM (accuracy = 71%). The lowest accuracy among the four algorithms was for algorithm No. 5, only 39%. Although the comparison results confirmed the superiority of the DM-IFSC, the accuracy rate is considered to be low. The nature of the dataset and the amount of overlap between samples have greatly affected the accuracy rate. Figure 4 shows a drawing of the straight lines of the dataset classes after calculating the linear equations where we notice that the lines are very close together. Despite the very close distance between the lines that makes it difficult to classify the unseen samples, the DM-IFSC produced 72% of classification accuracy.

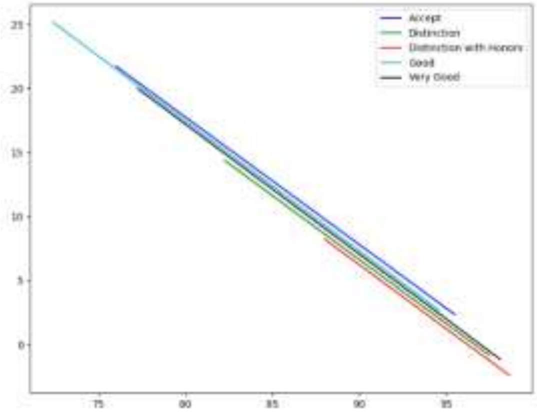

Figure 4 The straight lines of the dataset classes

By comparing the classification accuracy results presented in Table 11 with the results in Table 6, we can see that the efficiency of the DM-IFSC in solving binary classification problems is better than its efficiency in solving multiclass classification problems. Reducing the categories from 5 to 2 is the main reason for this comparative advantage in efficiency as this resulted in a somewhat spacing between the straight lines of the classes, thus improving the classification accuracy rate. However, this case s specific to the dataset that was used in this research paper. Figure 5 shows the straight lines of the two classes, "Accept" and "Distinction", and the distance between the two lines.

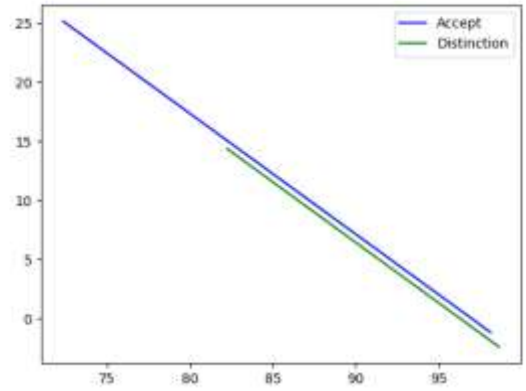

Figure 5 The straight lines of the two classes, "Accept" and "Distinction"

The results of experiments related to solving the multiclass classification problem and solving the binary classification problem indicate that the DM-IFSC is a good competitor to one of the most important classification algorithms, which is SVM. The DM-IFSC accuracy exceeded the accuracy of the SVM in the multiclass classification experiments, and its accuracy was close to its accuracy in the binary classification experiments. In terms of precision and recall, the DM-IFSC performed at higher level of precision (95%) than SVM, which means it outperforms the SVM and the other algorithms in terms of returning more relevant results than irrelevant ones. As for the recall values that can be considered as a measure of quantity, the DM-IFSC came in second place after the SVM algorithm, which came first. This confirms that the DM-ISFC returns most of the relevant results in binary classification.

6. Conclusion

Since the problem of non-linearly separable data points is common in many datasets, the researchers focused on finding a way around this problem and classifying the data with high accuracy. In this paper, a new IFS-based classification approach, namely DM-IFSC, is proposed and evaluated its performance on a highly overlapping educational dataset that is difficult to classify into a set of classes. The dataset is taken from records of 124 students who completed their bachelor degree. Each student is represented using 9 features and 1 class. The values of the 9 features refer to the student's scores in eight of the high school subjects he obtained in the 12th grade. The seventh feature is the current cumulative grade point average (CGA), and its value is from 0 to 4. The class of student represents his graduation grade. Firstly, the dataset is normalized to get an acceptable input dataset form that could be used to train and test the DM-IFSC. Secondly, the IFSs are constructed by considering the student's score as the membership degree $\mu_A(x)$, the student CGA as the $\pi_A(x)$, and the non-membership degree $\nu_A(x) \ = 1 - (\mu_A(x) + \pi_A(x))$. The IFSs created were used in modeling the classes of the dataset. The modeling task was carried out by calculating the equation of the straight line (slope) passing through the class IFSs. Lastly, the classification process is performed by calculating the distance between each class and the unseen sample that is subject to classifying. Comparisons of the classification performance of the DM-IFSC with three of the most common classification algorithms on the same dataset reveal the competitiveness and superiority of the proposed approach. The DM-IFSC achieved good results in binary classification problems, while its experiment results in solving the multiclass classification problem showed that the DM-ISFC approach needs further research. The method of constructing the IFSs represents another future direction. The performance of the proposed algorithm can be improved by searching for new methods of IFSs constructions especially in multiclass classification problems.

With regrads to research limitations, this study does not address the comuptational efficiency of DM-IFSC, which my be a concern for large-scale problems. Future unvestigations focused on optimizing the algorithm to reduce computational complexity could be condcuted. Another limitation is its ability to handle multiclass classification problems and improve overall performance. The integration of DM-IFSC with machine learning techniques, i.e. deep learning could be considered as an option for future research.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abdulla, S. and Al-Nassiri, A. (2015). kEFCM: kNN-based dynamic evolving fuzzy clustering method, in Proc. IJACSA. Citeseer. 5–13.

[2] Ahmad, B., Jian, W. and Anwar Ali, Z. (2018) Role of machine learning and data mining in internet security: standing state with future directions, *Journal of Computer Networks and Communications*, 2018.

[3] Alloghani, M. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science, Supervised and unsupervised learning for data science. 3–21.

[4] Atanassov, K. T. (1986). Intuitionistic fuzzy sets, *Fuzzy Sets and Systems, 20*(1). 87–96. doi: https://doi.org/10.1016/S0165-0114(86)80034-3.

[5] Aydilek, Ï. B. (2018). Examining effects of the support vector machines kernel types on biomedical data classification, in 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE 1–4.

[6] Bakhshinategh, B. (2018). Educational data mining applications and tasks: A survey of the last 10 years, *Education and Information Technologies, 23*(1). 537–553.

[7] Cai, Y. (2019). New classification technique: fuzzy oblique decision tree, Transactions of the Institute of Measurement and Control, 41(8), pp. 2185–2195.

[8] Chaira, T. (2019). Application of fuzzy/intuitionistic fuzzy set in image processing, Fuzzy Set and Its Extension: The Intuitionistic Fuzzy Set. 237–257.

[9] Dai, B., Wang, F. and Chang, Y. (2022). Multi-objective economic load dispatch method based on data mining technology for large coal-fired power plants, *Control Engineering Practice, 121*. 105018.

[10] Deborah, L. J. (2015). Fuzzy-logic based learning style prediction in e-learning using web interface information, Sadhana, 40(2), pp. 379–394.

[11] Ieracitano, C. (2022). A Fuzzy-enhanced Deep Learning Approach for Early Detection of Covid-19 Pneumonia from Portable Chest X-Ray Images', Neurocomputing.

[12] Islam, M. S. (2018). A systematic review on healthcare analytics: application and theoretical perspective of data mining, in Healthcare. MDPI 54.

[13] Jimenez-Carvelo, A. M and Cuadros-Rodríguez, L. (2021). Data mining/machine learning methods in foodomics, Current Opinion in Food Science, 37. 76–82.

[14] Kunnathuvalappil-Hariharan, N. (2018). Applications of Data Mining in Finance, Naveen Kunnathuvalappil Hariharan.(2018). APPLICATIONS OF DATA MINING IN FINANCE. *International Journal of Innovations in Engineering Research and Technology, 5*(2). 72–77.

[15] Meena, K. and Thomas, K. V (2018). An application of intuitionistic fuzzy sets in choice of discipline of study', Global J. Pure Appl. Math, 14(6). 867–871.

[16] Murthy, M. Y. B., Koteswararao, A. and Babu, M. S. (2021). Adaptive fuzzy deformable fusion and optimized CNN with ensemble classification for automated brain tumor diagnosis, *Biomedical Engineering Letters*. 1–22.

[17] Navas de Maya, B. (2022). Application of data-mining techniques to predict and rank maritime non-conformities in tanker shipping companies using accident inspection reports, *Ships and Offshore Structures, 17*(3). 687–694.

[18] Nguyen, G. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey, *Artificial Intelligence Review, 52*(1). 77–124.

[19] Peker, M. (2016). A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM, *Journal of medical systems, 40*(5). 1–16.

[20] Raja, J. B. and Pandian, S. C. (2020). PSO-FCM based data mining model to predict diabetic disease', Computer Methods and Programs in Biomedicine, 196. 105659.

[21] Ren, Q. (2022). A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree', *Journal of Petroleum Science and Engineering, 208*. 109681.

[22] Ricciardi, C. (2020). Application of data mining in a cohort of Italian subjects undergoing myocardial perfusion imaging at an academic medical center, Computer methods and programs in biomedicine, 189. 105343.

[23] Salem, H. (2022). Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes, Applied Sciences, 12(3) 950.

[24] Shojae-Chaeikar, S. (2020). PFW: polygonal fuzzy weighted—an SVM kernel for the classification of overlapping data groups, Electronics, 9(4). 615.

[25] Shubair, A., Ramadass, S. and Altyeb, A. A. (2014). kENFIS: kNN-based evolving neuro-fuzzy inference system for computer worms detection, *Journal of Intelligent & Fuzzy Systems, 26(*4). 1893–1908.

[26] Subhashini, L. (2022). Integration of fuzzy logic and a convolutional neural network in three-way decision-making. Expert Systems with Applications, 202. 117103.

[27] Thakare, K. V. (2022). A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection, *Expert Systems with Applications, 201*. 117030.

[28] Tugrul, F., Gezercan, M. and Citil, M. (2017). Application of intuitionistic fuzzy sets in high school determination via normalized Euclidean distance method, *Notes on Intuitionistic Fuzzy Sets, 23*(1). 42–47.

[29] Wu, X., Song, Y. and Wang, Y. (2021). Distance-Based Knowledge Measure for Intuitionistic Fuzzy Sets with Its Application in Decision Making, *Entropy, 23*(9). 1119.

[30] Xie, J. (2021). Network Intrusion Detection Based on Dynamic Intuitionistic Fuzzy Sets, IEEE Transactions on Fuzzy Systems.

[31] Xu, Y. and Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning, *Journal of analysis and testing, 2*(3). 249–262.

[32] Yan, H. (2020). Data mining in the construction industry: Present status, opportunities, and future trends, Automation in Construction, 119. 103331. doi: https://doi.org/10.1016/j.autcon.2020.103331.

[33] Zadeh, L. A. (1965). Information and control, *Fuzzy sets, 8*(3). 338–353.

[34] Zhang, H. (2014). Linguistic Intuitionistic Fuzzy Sets and Application in MAGDM, *Journal of Applied Mathematics. Edited by R. M. Palhares,* 432092. doi: 10.1155/2014/432092.