An Implementation of Intuitionistic Fuzzy Soft Set's Similarity Measure centered on Speech Emotion Recognition Distance

S. Mythili[1] and M.Pavithra[2*]

[1,2] Department of Mathematics, Excel Engineering Collage, Namakkal, Tamilnadu, India.

[*] *Corresponding author Email*: *pavithravcw.92@gmail.com*

Abstract

Accurately representing and analyzing large data sets is becoming more and more important in domains like as economics, pattern identification, medical diagnosis, and stock market analysis. As digital technology has spread, digitalized patterns and images have proliferated. These patterns are multidimensional, containing properties that are both physical and non-physical. A strong fuzzy-based model is required in order to represent and interpret this data efficiently. The Complex Vague Soft Set (CVSS) model is presented in this study. It is intended to concisely represent the multi-dimensional information included in digital photographs. In order to tackle pattern identification problems in digital images, this defines information measures for CVSSs, mainly concentrating on distance and similarity measurements. The study examines the connections between this similarity measure and the associated distance and provides an axiomatic definition of a distance-based similarity measure for CVSSs. Within the scope of CVSS, these linkages are both proposed and validated. This approach's usefulness is illustrated by applying it to a pattern recognition task, in which digital photographs are examined using multi-dimensional data, such as physical characteristics and extra metadata like timestamps and locations. This study demonstrates how well the CVSS model handles complicated data, making it a potent tool for raising the accuracy of pattern recognition in a range of applications.

Keywords: Hard to define fuzzy soft set - Similarity metric , Digital picture , Pattern recognition , Complex fuzzy set

1 Introduction

Uncertainty constitutes a fundamental element of the decision-making framework, thereby complicating the efforts of decision-makers to establish an effective systemInitially, it was posited that all relevant data regarding alternatives or objects was based on probabilistic and binary state models. Nonetheless, in real-world applications, these theoretical constructs frequently fail to manage information accurately, resulting in a lack of consistent beneficial insights for practitioners. To mitigate such difficulties, fuzzy set theory and its various extensions, including Pythagorean fuzzy sets, intuitionistic fuzzy sets, and neutrosophic sets, are essential for addressing uncertainties within data by allocating membership degrees to individual elements. Building on these foundational principles, a multitude of aggregation operators have been proposed by various researchers, including Aggregation methodologies, including mean calculations, geometric aggregation techniques, alongside other related operators (ALBAITY et al, 2023). From the analyses presented in the existing literature, it has been observed that these investigations have examined decision-making scenarios through the lenses of fuzzy sets, intuitionistic fuzzy sets, or their generalizations, which are primarily capable of tackling the uncertainties and ambiguities inherent in the data. Essentially, Scholars across various academic fields are increasingly focusing on the procedure of constructing and representing data characterized by fuzzy attributes. Within this domain, two primary challenges arise: (1) The presence of intricate data sets that exhibit maximal uncertainty and vagueness, often comprising unreliable and incomplete information; (2The characteristics and constituents of these data sets exhibit an intrinsically two-dimensional quality, which clarifies the extent to which elements are interconnected with a defined array of attributes and how these attributes experience temporal transformation. (GARG, 2017) The prevailing models fail to adequately encompass the partial ambiguity linked to the data and its variability at any given moment. However, within complex data sets, the

simultaneous occurrence of ambiguity and uncertainty, along with variations in the data's periodic phase, warrants particular attention.

(GARG , 2016) Complex data sets are commonly encountered in fields such as medical research and government databases, where they encompass diverse types of information, including biometric data, facial recognition, audio recordings, and images. These datasets are often characterized by significant amounts of incomplete, uncertain, and ambiguous data. With the increasing prevalence of big data analytics, effectively managing these intricate and voluminous data sets has become a critical challenge for researchers.

CFS integrates fuzzy set theory with complex numbers, offering a way to manage the cyclical characteristics of two-dimensional data sets. However, CFS has limitations: (1) it enforces a strict membership framework that requires a single value for membership, overlooking the inherent partial uncertainties in the data, and (2) its parameterization is somewhat limited (GARG et al., 2018).

(GHOSH et al., 2023) The procedure of conceptualizing and representing data characterized by fuzzy attributes (CIFS) was developed. CIFS improves upon CFS by introducing a non-membership degree and more clearly defining the operations of union, intersection, and complement. Alkouri and Salleh later expanded this by introducing a distance metric for CIFS that incorporates compositions, projections, and complex intuitionistic fuzzy relations. They also explored various metrics for fuzzy soft sets with complex intuitionistic properties. These advancements have employed in the realm of medical diagnostics and pattern recognition, thereby illustrating the significance of CIFS in handling complex, uncertain data (KHAN et al., 2019).

In the domain of similarity (or distance) metrics, a variety of measures for imprecise sets have been introduced. For these ambiguous constructs, numerous entropy, similarity, and distance measures have been formulated. Additionally, distance metrics have been proposed specifically for type-2 fuzzy intuitionistic sets. (MUHSEN et al., 2023)The application of set-theoretic methodologies, the utilization of matching functions, and the employment of the geometric distance framework have all been instrumental in the formulation of similarity metrics for Intuitionistic Fuzzy Sets (IFSs), which have since been utilized to tackle decision-making challenges. Within the framework of IFS and interval-valued IFS contexts, a connection number grounded in set pair analysis theory has been devised utilizing the TOPSIS methodology to resolve decision-making complexities (QIUand LI, 2017).

Moreover, analyses of similarity and distance concerning the intuitionistic multiplicative preference relation have been elucidated, highlighting their significant relevance in decision-making contexts. Innovative similarity metrics because intuitionistic fuzzy sets, commonly referred to as IFSs, possess introduced and utilized to tackle challenges in pattern recognition. Within the IFS framework, a generalized entropy metric of specified order and extent has been proposed and employed in decision-making contexts. Multiple Similarity measures for fuzzy soft sets exhibiting generalized intuitionistic characteristics within the scope of soft sets, applying these metrics to resolve issues in texture synthesis and medical diagnostics. Regarding dual hesitant fuzzy soft sets, has introduced an array of distance and similarity metrics that have found applications in pattern recognition and decision-making processes (SALEEM et al., 2023)

Additionally, researchers have explored various informational metrics within the fields of fuzzy soft sets, ambiguous soft sets, fuzzy sets, and complex sets, as outlined in references. Contemporary methods, as previously noted, employ fuzzy sets and hybrid models based on fuzzy logic, including intuitionistic fuzzy sets and soft sets, to tackle issues in fuzzy pattern recognition. Images (restricted to analog formats) are represented through fuzzy sets, and fuzzy information metrics are utilized to assess divergence, similarity, and entropy, thereby enhancing process capabilities. However, owing to the limitation that these techniques can manage only a single dimension of data at any given time, their applicability is confined to resolving pattern recognition issues involving straightforward analog images (SIDIROPOULOS, 2022).

(SINGH and GARG 2017) Currently, there is a notable gap in the scholarly literature regarding fuzzy pattern recognition methods that can effectively handle the recognition of digital images with multi-dimensional

characteristics datasets. These datasets encompass not only the two-dimensional data related to tangible features of images—such as contours, edges, rotations, and pixel dimensions—but also intangible aspects like timestamps and geographic coordinates. These diverse characteristics are fundamental to digital images and their patterns. Additionally, a crucial factor to consider is the partial ignorance associated with the data, which may stem from the inherent uncertainty and imprecision within the information.

The combination of attributes from Complex Fuzzy Sets (CFSs), soft sets, and vague sets has led to the creation of the theory of Complex Vague Soft Sets (CVSS). This advanced framework provides an improved alternative to traditional fuzzy sets for managing two-dimensional datasets, effectively addressing the challenges previously identified. By tackling these challenges and presenting an optimal substitute for CFSs within the domain of two-dimensional information processing, the CVSS framework emerges as a more comprehensive and robust evolution of the CFS model (YAQOOT et al., 2023).

(Yu et al., 2022) The CVSS framework demonstrates a distinct advantage over CFSs and other intricate fuzzy-based systems due to its unique structural characteristics, which include:

(1) **Interval-Based Membership**: Users can articulate their uncertainty by using intervals to assign membership values to various components.
(2) **Handling Partial Ignorance**: CVSS adeptly manages the partial ignorance in data, accommodating inherent uncertainty and imprecision.
(3) **Advanced Parameterization**: The framework includes sophisticated parameterization features that refine the representation of parameters, enhancing overall precision.

Motivated by the attributes of the CVSS framework, this research seeks to explore the compositional attributes of CVSSs alongside their descriptive metrics pertinent to the administration of multi-dimensional intricate datasets encompassing digital imagery and patterns. An axiomatic formulation of the similarity metric among CVSSs is developed utilizing the characteristics of CVSSs, CFSs, and fuzzy sets. This formulation is then used to develop a novel function for assessing the similarity between two Complex Vague Soft Sets (CVSSs). Additionally, considerable research has been dedicated to tackling the difficulties of pattern recognition in digital images and in multi-dimensional complex datasets.

This study will concentrate regarding the discernment and depiction of digital imagery utilizing multi-dimensional intricate datasets, employing the attributes of Complex Valued Soft Sets (CVSSs) and a newly developed similarity metric specifically designed for this framework, thereby enhancing the clarity of the proposed methodology. The outcomes and analyses generated by the recommended approach are compared with those obtained from a variety of alternative models and their associated methodologies to substantiate the conclusions reached. The subsequent sections of this scholarly document are carefully structured as outlined below. Section 2 offers a comprehensive overview of key concepts related to complex fuzzy sets and other foundational principles that underpin this study.

In Section 3, the axiomatic characterization of the similarity metric is articulated and substantiated, alongside a discussion of its fundamental attributes. The applicability of the CVSS paradigm and its similarity metric is illustrated in Section 4 through an example involving digital images characterized by multi-dimensional data for the purpose of pattern recognition. In Section 5, we conduct a succinct comparative evaluation between the methodology delineated in this investigation and other pertinent strategies identified in the existing literature, aiming to underscore the superiority and practical significance of our proposed approach. The final section presents concluding reflections.

Evaluation Protocol

This framework encompasses a systematic approach to pattern identification, facilitating the extraction of IFS patterns pertinent to each unique document type. By leveraging these the classification technique enables the systematic categorization of any newly introduced document by recognizing its corresponding IFS patterns. The

approach to classification involves ascertaining which category the IFS patterns bear the closest resemblance to or the greatest divergence from the newly presented document.

The initial phase of the framework, which integrates document preprocessing with IFS pattern extraction, is illustrated in Figure 1. This phase encompasses the refinement of the source literature, transforming them into a bag-of-words representation, and reorganizing them to highlight IFS. Ultimately, pattern extraction—referred to as IFS-centric representation is employed to depict the IFS patterns that are relevant to each specific documentclassification.

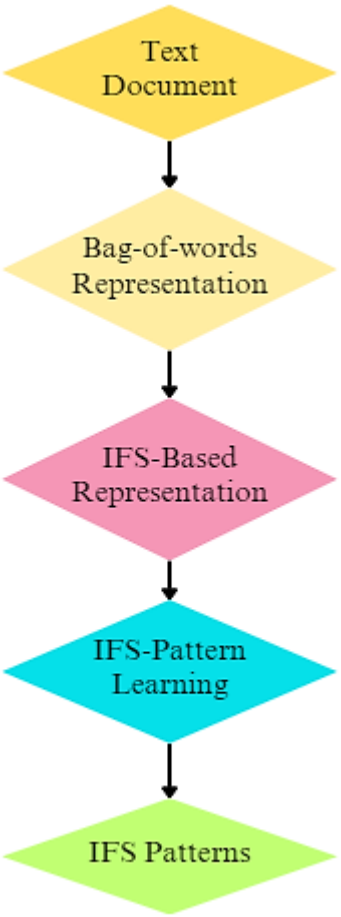

Figure 1. The study's document learning protocol was used.

During the evaluative phase, a test document di undergoes a conversion into its IFS representation, and the classification of the document is established by identifying the class Ck that most closely aligns with the test document concerning its pattern. A distance or similarity metric is employed to ascertain the nearest pattern.

2.Bag-of-Words Representation

This exposition delineates the prevalent methodology for converting textual documents into a quantitative format within the domain of Natural Language Processing (NLP). By utilizing the frequency of occurrence of each term, this technique constructs a representation of a textual document. To formulate a bag-of-words, it is necessary to either establish a lexicon of the recognized terms (the vocabulary to be incorporated) or to implement a threshold criterion to eliminate terms that appear with low frequency. Consequently, for h terms present in a document, the bag-of-words vector V can be mathematically articulated as follows (eq1):

$$V = (n_{i,1}, n_{i,2}, n_{i,3}, \ldots, n_{i,h}) \quad (1)$$

where n is the number of occurrences of each word of the vocabulary.

2.1 The IFS Depiction

The subsequent methodologies were employed to transform a bag-of-words depiction of a document di into a representation grounded in Intuitionistic Fuzzy Sets (IFS). It is imperative to elucidate in eq 2:

$$zi,j = \frac{n_{i,j} - mean_j}{std_j} \quad (2)$$

where meanj\text{mean}_jmeanj and stdj\text{std}_jstdj denote the arithmetic mean and standard deviation, respectively, of the occurrences of word jjj across all documents in each class, while ni,jn_{i,j}ni,j represents the frequency of word jjj in document did_idi. Accordingly, a weighted sigmoid function is employed to determine the membership and non-membership values for did_idi.

$$\mu_{i,j} = \frac{r_j}{1 + e^{-zi,j}} \quad (3)$$

$$v_{i,j} = \frac{r*j}{1 + ezi,j} \quad (4)$$

where rjr_jrj and rj∗r^∗_jrj∗ represent weights within the range [0, 1]. The degree of hesitation is subsequently determined using Equation (3 and 4).

2.3IFS Pattern Learning

Finally, in the pattern learning phase, the IFS pattern for class CkC_kCk, denoted as PkP_kPk, is defined as follows (eq 5):

$$P_k = \{(\tilde{\mu}_k, \tilde{v}_k)\} \quad (5)$$

where $\tilde{\mu}_k$and $\tilde{v}_k$represent the average membership and non-membership values of all documents within the class C_k.

2.4 Document Classification

Document categorization relies on the extent to which the characteristics of a document diverge (in terms of distances) or align (regarding similarities) with the established class patterns identified in the preceding phase. In the process of learning class patterns, the meanj and stdj values are employed to derive the IFS representation for an incoming document (eq 6 and 7).

To classify a new document into class C′C'C′, the following equations can be applied to determine its class:

$$C' = \text{argmin}\{\text{Sim}(Pk, IFSdn)\} \quad (6)$$

where Sim(Pk,IFSdn)\text{Sim}(P_k, \text{IFS}_{dn})Sim(Pk,IFSdn) represents a similarity measure between PkP_kPk and IFSdn\text{IFS}_{dn}IFSdn.

$$C' = arg\ \max\{Sim(P_k, IFS_dn)\} \quad (7)$$

where Pk denotes the class structures established for the classification Ck throughout the IFS pattern acquisition stage; IFSdn denotes the IFS characterization pertaining to the new document dn; and Sim and Dist represent the respective metrics of similarity and distance, that are employed to evaluate the IFSs.*Experimental Setup*

2.5 Measures of Similarity and Distance

The first column of The names of the authors are included in Tables 1 who originally introduced each metric, while the second column specifies the parameters required for the functions used to compute the corresponding metric (excluding the resemblance nature), along with the two Intuitionistic Fuzzy Sets, labeled as A and B. The function, detailed in the third column, was first introduced by the authors in their foundational work to mathematically define the metric.

The designation has been revised to dh2, as two metrics presented by under a similar designation were identified as dh. Furthermore, Appendix A encompasses the mathematical formulations of all the measures.

2.6 Datasets

The subsequent section enumerates and elucidates the datasets employed in our experimental procedures:

- BBC News: The dataset presented encompasses 2225 entries from the years 2004 and 2005, which have been systematically classified into five distinct thematic domains (business, entertainment, politics, sports, and technology), amounting to 9635 words in total.
- BBC Sports: This dataset includes 737 articles categorized into five distinct sports genres— athletics, cricket, soccer, rugby, and tennis —totaling 4,613 words, similar to the dataset mentioned earlier. Notably, the BBC News dataset shows a balanced distribution, with 386 documents classified under entertainment and 511 documents categorized as sports. Furthermore, the word count within the classes demonstrates balance, with the business and technology classes exhibiting the highest and lowest average word counts, respectively, at 513 and 339.

Table 1: Average Word Count and Class Distribution in BBC Sports and News Datasets

Classes	Number of samples	Mean number of words per text
Mean Business	500	350
Entertainment	380	350
Politics	420	450
Sport	500	350
Tech	400	500
Athletics	100	220
Cricket	150	260
Football	250	240
Rughy	150	250
Tennis	100	200

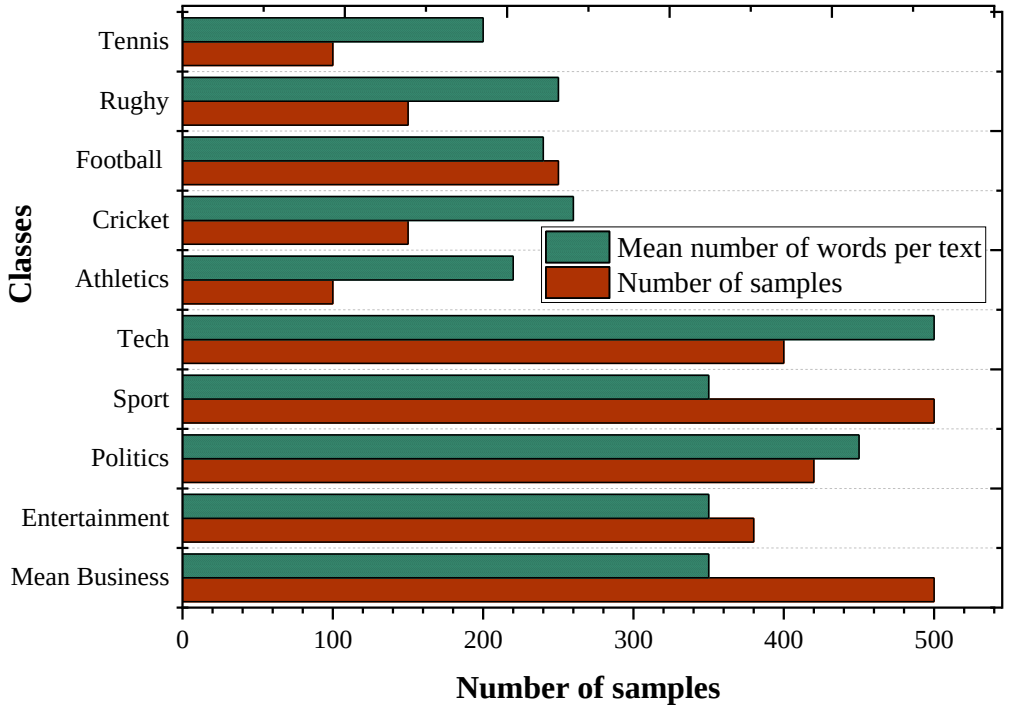

Figure 2. The datasets derived from BBC Sports and News illustrate the average word count for each sample text in conjunction with the distribution of classes.

The distribution of samples reveals a significant disparity within the BBC Sports dataset. The football category possesses the most substantial sample size (265). In contrast, the average word count per text remains relatively consistent across all categories, with the exception of tennis, which has the lowest mean word count (298), while the cricket category records the highest (391).

2.6.1 Information Gathering

Lemmatization is the process of reducing various forms of a word to its base or root form. For example, "does," "doing," and "did" are all converted to "do," while "am," "are," and "is" are reduced to "be." This procedure was executed using the Scikit-learn and Natural Language Toolkit libraries. Furthermore, the BBC News dataset produced a vocabulary of 2,540 words, while the BBC Sports dataset contained 2,396 words, which reflects the cut-off frequency selected through a trial-and-error approach.

2.7 Assessment of Performance

1 The effectiveness of each metric was evaluated through accuracy, precision, recall, and F1-score, in addition to the Degree of Confidence (DoC). The fsmpy library was used to compute the Degree of Confidence, while the remaining metrics were determined using the Scikit-learn library. These metrics are widely used in classification studies to assess the performance of algorithms or models, with the DoC frequently cited in academic discussions on feature selection (FS) theory.

2 The Degree of Confidence (DoC) rigorously evaluates the extent to which the metric reliably categorizes a specific sample as conforming to an established framework. Therefore, elevated DoC values signify that the metric exhibits a higher degree of certainty in the classification of a sample. Additionally, this employed their "macro" averaging methodology to ascertain the precision, recall, and F1-score. This process entailed computing these metrics for each label individually and then averaging them to derive their unweighted mean.

3. Results

In this study, this evaluated the performance of 19 distance metrics and 43 similarity metrics for classifying textual documents from two separate datasets. The analysis was performed using Python, with the Scikit-learn and fsmpy libraries. A grid search strategy was utilized to test various parameter combinations rjr_jrj and rj∗r^*_jrj∗, specifically five values evenly spaced within the range [0.1, 1]: 0.1, 0.325, 0.55, 0.775, and 1.

After identifying the optimal parameters and corresponding membership and non-membership weights for each metric, this applied a leave-one-out cross-validation approach to assess their performance across all datasets. In this method, one document was held out for testing in each iteration.

5. Discussion

The aforementioned comparative evaluations substantiate that the results obtained from our proposed methodology are consistent with those produced by other analogous and prevalent techniques within this field. Although these alternative methodologies indicate that various hybrid models of fuzzy sets may also achieve similar outcomes, they further suggest that the decision-making framework is significantly more intricate and demands a greater amount of computation, rendering it less computationally efficient than our approach recommended approach.

In contrast, our proposed model succinctly encapsulates the information related to both the tangible and intangible aspects of images within a singular framework. As a result, this can directly compute the similarity among images utilizing our suggested methodology without the necessity to allocate weights to the attributes. This approach negates the requirement for supplementary operations to amalgamate the data concerning the tangible and intangible aspects of the images. When evaluated against the other methodologies discussed, the capability of our proposed technique to produce equivalent outcomes while demanding less computational resources serves as evidence of its enhanced computational efficacy.

The validation of our proposed methodology reinforces its feasibility and efficiency in addressing intricate data sets and multifaceted pattern recognition challenges.

6 Conclusion

In this research paper, this have established an axiomatic framework for the distance-based similarity metric within the CVSS model and scrutinized several critical characteristics of this metric. Furthermore, the interrelations between the similarity and distance metrics pertinent to CVSS were elucidated and substantiated. These interconnections yielded innovative formulas, derived from the CVSS distance metrics, for the comparative analysis of similarity among CVSS instances. A concise case study was conducted, concentrating on the application of multi-dimensional datasets encapsulated within images and patterns, resulting from the digital transformation of these entities. To highlight the importance and effectiveness of the CVSS model and its similarity metric, this presented an illustrative example related to an efficient pattern recognition challenge utilizing multi-dimensional datasets to characterize digital imagery. In conclusion, this executed an exhaustive comparative analysis between the methodologies delineated herein and other analogous techniques identified within the extant literature, with the intent to substantiate and exhibit the predominance of our proposed model and its corresponding methodology concerning computational efficiency, practicality, and the representational clarity of information. The findings presented in this manuscript may be further developed in subsequent research to include multiplicative preference frameworks, complex neutrosophic sets, and various ambiguous fuzzy contexts.

References

1. ALBAITY, M., MAHMOOD, T., and ALI, Z., 2023, Impact of Machine Learning and Artificial Intelligence in Business Based on Intuitionistic Fuzzy Soft WASPAS Method, Mathematics, 11(6):1453.

2. GARG, H., 2016, A Novel Correlation Coefficients Between Pythagorean Fuzzy Sets and Its Applications to Decision-Making Processes, International Journal of Intelligent Systems, 31(12):1234-1252.
3. GARG, H., 2018, Some Methods for Strategic Decision-Making Problems With Immediate Probabilities in Pythagorean Fuzzy Environment, International Journal of Intelligent Systems, 33(4):687-712.
4. GARG, H., AGARWAL, N., and TRIPATHI, A., 2017, Generalized Intuitionistic Fuzzy Entropy Measure of Order α and Degree β and Its Applications to Multi-Criteria Decision Making Problem, International Journal of Fuzzy System Applications (IJFSA), 6(1):86-107.
5. GHOSH, S. K., GHOSH, A., and BHATTACHARYYA, S., 2022, Recognition of Cancer Mediating Biomarkers Using Rough Approximations Enabled Intuitionistic Fuzzy Soft Sets Based Similarity Measure, Applied Soft Computing, 124:109052.
6. KHAN, S., ABDULLAH, S., and ASHRAF, S., 2019, Picture Fuzzy Aggregation Information Based on Einstein Operations and Their Application in Decision Making, Mathematical Sciences, 13:213-229.
7. MUHSEN, Y. R., HUSIN, N. A., ZOLKEPLI, M. B., and MANSHOR, N., 2023, A Systematic Literature Review of Fuzzy-Weighted Zero-Inconsistency and Fuzzy-Decision-By-Opinion-Score-Methods: Assessment of the Past to Inform the Future, Journal of Intelligent & Fuzzy Systems, 45(3):4617-4638.
8. QIU, J., and LI, L., 2017, A New Approach for Multiple Attribute Group Decision Making With Interval-Valued Intuitionistic Fuzzy Information, Applied Soft Computing, 61:111-121.
9. SALEEM, N., KHATTAK, M. I., ALQAHTANI, S. A., JAN, A., HUSSAIN, I., KHAN, M. N., and DAHSHAN, M., 2023, U-Shaped Low-Complexity Type-2 Fuzzy LSTM Neural Network for Speech Enhancement, IEEE Access, 11:20814-20826.
10. SIDIROPOULOS, G. K., DIAMIANOS, N., APOSTOLIDIS, K. D., and PAPAKOSTAS, G. A., 2022, Text Classification Using Intuitionistic Fuzzy Set Measures—An Evaluation Study, Information, 13(5):235.
11. SINGH, S., and GARG, H., 2017, Distance Measures Between Type-2 Intuitionistic Fuzzy Sets and Their Application to Multicriteria Decision-Making Process, Applied Intelligence, 46:788-799.
12. YAQOOT, I., RIAZ, M., and AL-QURAN, A., 2023, New Similarity Measures and TOPSIS Method for Multi Stage Decision Analysis With Cubic Intuitionistic Fuzzy Information, Journal of Intelligent & Fuzzy Systems, (Preprint):1-24.
13. YU, D., SHENG, L., and XU, Z., 2022, Analysis of Evolutionary Process in Intuitionistic Fuzzy Set Theory: A Dynamic Perspective, Information Sciences, 601:175-188.