https://doi.org/ 10.33472/AFJBS.6.9.2024.438-452

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Entropy assisted intuitionistic fuzzy rough bireduct method to improve the prediction of induced peculiar peptides

Aneesh Kumar Mishra
Department of Computer Science Engineering
Jaypee University of Engineering & Technology Guna (M.P.),India
aneeshjuetphdcse2k19@gmail.com

Neelesh Kumar Jain
Department of Computer Science Engineering
Jaypee University of Engineering & Technology Guna(M.P.),India
neelesh.dei@gmail.com

Ravindra Kumar Singh
Department of Electronics & Communication Engineering
Christ University, Bangalore
singh.ravindrakumar@gmail.com

Article History

Volume 6,Issue 9, 2024

Received:17 Mar 2024

Accepted : 11 Apr 2024

doi: 10.33472/AFJBS.6.9.2024.438-452

Abstract

Rough set theory demonstrated as a successful tool for data reduction in the last few years. Fuzzy rough set discarded the limitations of rough set theory to handle real-valued datasets. Intuitionistic fuzzy rough set was incorporated as an interesting tool to cope with the vagueness and uncertainty in a much better way when compared to fuzzy set based approaches specifically in the scenario where uncertainty was inserted not only by judgement but also by identification. In this paper, we establish a new intuitionistic fuzzy rough set model based on the concept of entropy to deal with imbalanced dataset. This model is used to handle both data size and dimension simultaneously. Firstly, synthetic minimization optimization technique (SMOTE) is applied to induce relatively balanced datasets. Secondly, positive region constructed from the lower approximation of the proposed model is used to remove redundant size. Thirdly, this positive region is applied to establish a dependency function assisted technique of feature selection to eliminate redundant and irrelevant attributes/features. Next, various learning method is incorporated to investigate the performance measures over the produced reduced datasets. Then, induced peculiar peptides prediction is improved by using entire methodology. The best results are produced by support vector machine with RBF kernel with sensitivity, accuracy, specificity, MCC, and AUC as 87.9%, 87.7%, 84.0%, 0.697, and 0.898 respectively on the basis of 10-fold cross validation.

Keyword: Rough Set, Bireduct, Fuzzy set, SMOTE, SVM, Intuitionistic Fuzzy Set

1. Introduction

An interesting immune-regulatory cytokine is IL-13 [1], which is mainly excreted by stimulated T helper-Type 2 cells, and is used to inhibit inflammatory cytokine production. In the recent articles, it has been shown that IL-13 is also applied for the production of diverse cell types, basophils, including eosinophils, smooth muscle cells, mast cells, fibroblasts, and natural killer cells [2] with speckled biological functions [3]. GATA3 transcription factor is originally used to regulate the transcription of IL-13. Human chromosome 5q31 is surrounded by IL-13 with approximately 25% sequence homology along with IL-4. From the literature, it can be observed that IL-13 appears to be more encouraging target to design therapeutics rather than IL-4, whilst IL-4 and IL-13 are functionally highly related. IL-13 is implemented as an intermediate for numerous essential functions in various biological pathways comprising regulation of mastocytosis, airway hyper responsiveness, goblet cell hyperplasia, allergic inflammation, tumor cell growth, tissue eosinophilia, tissue remodeling, IgE Ab production, fibrosis, and intracellular parasitism. IL-13 also impedes very important tumor immuno-surveillance, which may control successfully to carcinogenesis [13, 16, 34]. In a fascinating study, different articles have discussed differential countenance levels of diverse 14 cytokines incorporating IL-13 in healthy regulator, moderate, and austere COVID-19 patients. Moreover, it has been outlined that the greater IL-13 expression level is found to be directly proportional to the severe COVID-19. Therefore, it is highly required to present a prediction methodology dedicated to perform an accurate classification of IL13 inducing peptides. From the previous investigation, it can be identified that very few techniques are given to discriminate the IL-13 inducing and non-inducing consisted peptides from available amino acid sequences.

In the current era of machine learning, data is increasing rapidly and becoming a very integral portion with the advancement of technologies [4]. Data is increasing both in terms of data points and dimensionality. So, an organized tool is always in demand to cope with this large volume of data [25, 31]. These data are usually consisted of uncertainty, noise, vagueness, imprecision, redundant instances, irrelevant and/or redundant features. In the recent years, various approaches are successfully presented and implemented to tackle these issues [28, 30]. These techniques are still facing well-known problems such as time, storage and cost. In the literature different Feature selection and instance selection techniques are available that can handle some of the above mentioned issues. Traditional feature and instance selection techniques were using the feedback from the experts, which were leading to major information loss [56].

Conventional rough set theory is initially proposed by the research article presented by Pawlak [35, 36]] as an interesting mathematical tool for not only data analysis but also for the knowledge discovery. Datasets with nominal features were effectively handled by this theory as a powerful concept in discovering knowledge [24, 33]. But, discretization leads to huge information destroy by the rough set based methodologies while handling real-valued datasets. As an improved extension of this traditional rough set theory, Dubois and Prade [14, 15] established fuzzy rough set theory, which was efficiently applied to manage the real-valued or even different mixed datasets. Fuzzy rough set aided techniques have been profitably applied in the areas of signal processing, data mining, bioinformatics, biomedical and machine learning. One of the major contributions of the fuzzy rough sets is effective data reduction including dimensionality reduction [7, 10, 37], rule extraction, instance selection [54, 61], decision making and so on. Discernibility matrix, fuzzy information-entropy [33,63], and fuzzy-dependency-function abetted techniques are the main notions of fuzzy-rough-set-facilitated dimensionality reduction methods [26, 27]. From the literature [40, 45, 46, 47, 48, 49, 50] it can be concluded that very few instance selection approaches are presented by using fuzzy rough set theory when compared to dimensionality reduction techniques [25, 29, 61]. In the preliminary work, membership to the computed fuzzy positive region was used to choose non-redundant instances, where this value was not less than an assumed pre-specified threshold. A prototype selection was presented by using fuzzy rough concept, where a wrapper method was given to determine the quality instances [60]. A weighted sampling idea to identify the representative samples by K-nearest neighbor based rule was introduced. Recently, fuzzy granules was incorporated by Zhang *et al.* [62] to design a fuzzy-rough-set to select representative instances as per the discriminating power of the fuzzy granules in a well discussed fuzzy decision system. Still, elimination of the noise from the available original datasets was not reviewed in this method, this may include the noise in the representative sample set. Moreover, computation time and complexity was handled by Zhang et al. [63] by introducing the incremental feature selection based on the representative examples in dynamic environment. Further, various research articles designed simultaneous attribute reduction and instance selection, where main focus was pointed out on the optimization of intelligent algorithms. Optimization of intelligent algorithms is a kind of approach, which simulates natural circumstances and behaviors with the existing population-assisted iterations [54]. In particular, Anaraki et al [64] discussed steady-state genetic and shuffled frog leaping algorithms for the simultaneous attribute and sample selection concepts. Furthermore, the notion of a bireduct was presented [23, 31], where an extension of the idea of a reduct based rough set theory was demonstrated. A bireduct comprises both some data points and features, and appears to be a class of rules of classification. Simultaneous dimensionality and instance reduction by using fuzzy rough set theory was investigated and analogous algorithms along with a frequency associated approach taking as a heuristic was established to choose the data points and/or features alternatively. Moreover, an interesting bireduct technique was designed with the help of fuzzy rough-set theory by using a harmony-search algorithm [32].

Rough set theory [35] was successfully applied to avoid the vagueness and expert feedback, but it was still facing the issues of information loss as it cannot be applied to the real-valued datasets without discretization [24, 33]. This problem was discarded with help of fuzzy rough set based techniques. Numerous extended fuzzy rough set models were established by the researchers to avoid all the above mentioned issues. But, fuzzy set theory [51, 59] fails to deal with the problems, where judgement and identification both are important. Intuitionistic fuzzy set [5, 6, 52] assisted techniques were introduced to tackle this failure.

In the current work, we have introduced an entropy-based intuitionistic fuzzy rough set [19, 20, 21] based methodology to enhance the discrimination of IL-13 inducing and non-inducing peptides. Firstly, an intuitionistic fuzzy entropy based tolerance relation is introduced. Secondly, lower and upper approximation is established based on this relation. Thirdly, sequences composed of IL-13 inducing and non-inducing peptides are taken from et al. [1]. Ifeature web server [11] is used to extract 953 features vectors of fixed length. Then, SMOTE (Synthetic minority optimization technique) [8] is applied to construct a balanced dataset. Next, lower approximation based positive region is presented to eliminate redundant samples. Moreover, this positive region is used to exhibit degree of dependency assisted attribute reduction. Further, irrelevant and redundant features are avoided by using this attribute reduction. Thereafter, several learning algorithms are investigated over this reduced dataset after applying above bireduct. An important schematic notation of entire methodology as discussed above is given in Figure 1.

2. Material and Methods

2.1. SMOTE

In this section, we illustrate the traditional SMOTE to cope with the class imbalance problem. The encountered problem of training with any machine learning technique over imbalanced datasets was degraded by the minority class, which was not capable to generate a promising performance on the minority class. The most straightforward technique was elaborated to deal with an imbalanced dataset by creating the required minority class samples, i.e. oversampling of the interesting minority class [12]. The informal approach is to repeat the available entries of the minority class. Here, the newly incorporated entries does not allow any valuable or novel information.

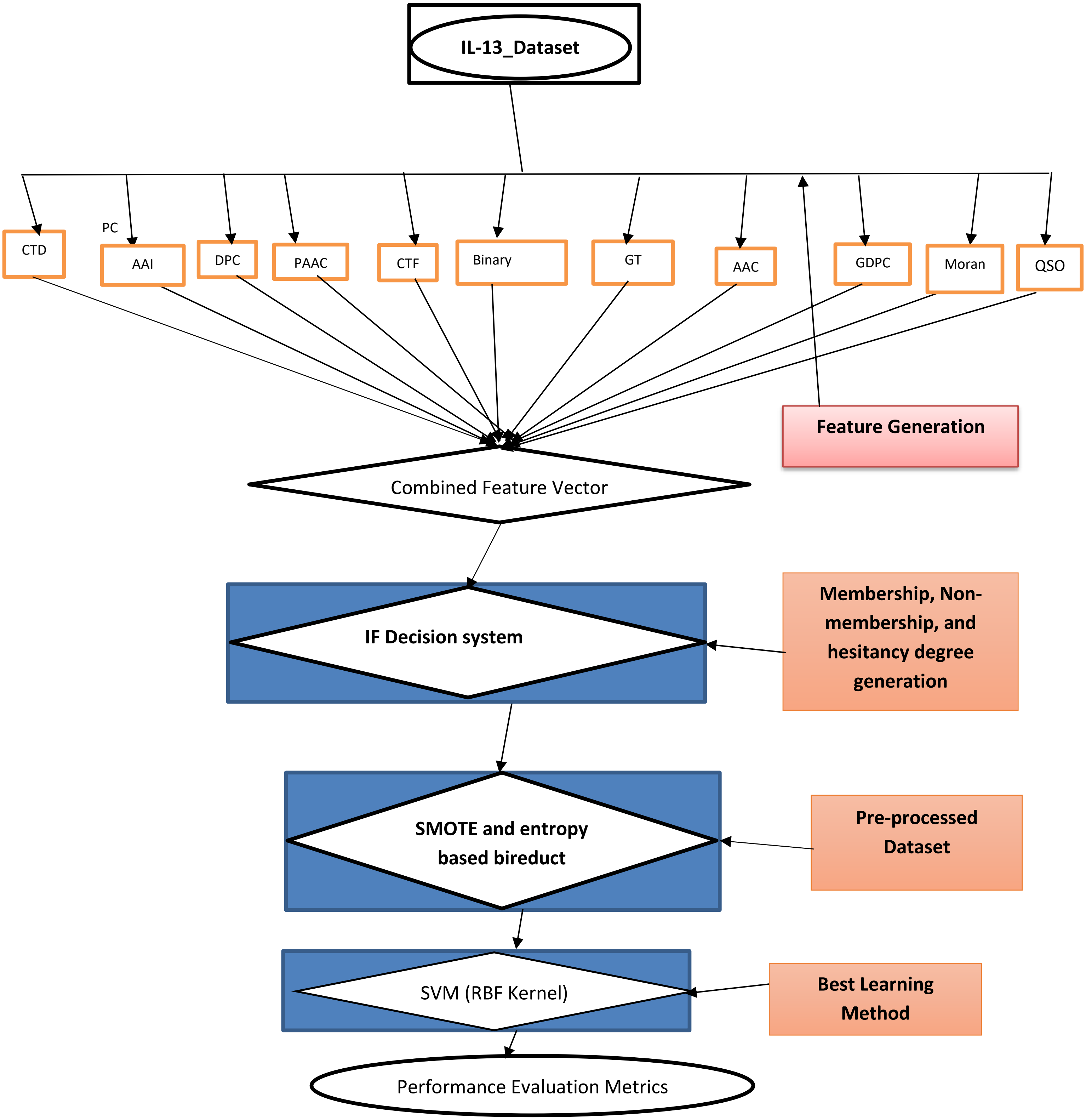

Fig. 1. Schematic depiction of the entire proposed methodology to enhance the prediction of induced peculiar peptides

On the other hand, new entries can be created by using the inventive minority class instances. It was effectively done with Synthetic Minority Oversampling Technique, which is abbreviated as SMOTE, and is very effective for existing tabular data. This concept of the data augmentation is found to be a much better improvement from restating/replicating entries from the instances related to minority class.

In unpretentious words the approach of SMOTE [53] chooses the data points which are closer to the accessible attribute space. Then, it tries to mark a best fit unavoidable line in between the samples of feature space. Thereafter, it appends a new sample for minority class instances by selecting artificial instances at a point on the effectively drawn line.

To be more precise, SMOTE chooses a random data point's entry from the overall obtainable minority class. Further, some k-nearest neighbors for that data point are computed. Next, from these k-neighbours a data point is selected randomly, and a sample is synthesized by randomly obtaining a data point on the drawn line between the data point and the chosen k-neighbour in the handy feature space. SMOTE includes synthesizing important entries of the minority class by selecting randomly to the minority class data point from its handy k-nearest neighbours. In addition with the above mentioned concept, data cleaning techniques are used to avoid the entries, which may mislead the entire classification process.

2.2. Dataset

We have adopted dataset from Jain et al. [1]. Here, 343 IL-13 inducing experimentally validated peptides of humans was extracted from already discussed IEDB. Throughout pre-processing task, the peptides consisting of length less than 8 or even more than 35 amino acids were eliminated. Moreover, duplicate numbers of the peptides were discarded. In the end, a positive dataset composed of 313 IL-13 inducing peptides were remained for next step. One of the main obstacles was to compile negative samples of the dataset to experimentally validate non-inducing peptides. To avoid this issue, the negative samples of the dataset was recorded from the recently published research article IL6Pred. Thereafter, the pre-processing was done to collect 2908 non-inducing peptides. Finally, a dataset was acquired with positive samples of 313 IL-13 inducing and a negative samples of 2908 non-inducing peculiar peptides.

2.3. Feature Generation

Peptide data points can be reformulated with abundant and suitable attribute vectors that replicates the necessary correlation along with the favorite target to advance a commanding predictor. In this study, the input attribute/feature vectors are generated by mining the following diverse attributes from distinct peptide sequences as discussed in the literatures [11, 22].

I. Amino acid composition (AAC)
II. Composition/Transition/Distribution (CTD)
III. Dipeptide composition(DPC)
IV. Binary Composition (Binary)
V. Pseudo-amino acid composition (PAAC)
VI. Moran correlation (Moran)
VII. Conjoint triad (CTF)
VIII. Grouped tripeptide composition (GTPC)
IX. Quasi-sequence order (QSO)
X. Grouped dipeptide composition (GDPC)
XI. Amino acid index (AAI)

Features are constructed of fixed length based on the information from iFeature web server. Overall 953 features are produced for each protein sequences either for the positive samples or for the negative samples.

3. Intuitionistic Fuzzy Entropy based bireduct method

Let us assume that Ꜧ depicts the collection of samples. Let $\underline{\text{Ḉ}}$ and $\underline{\text{Đ}}$ be non-empty determinate assemblage of conditional features and assemblage of decision features respectively, which forms an information scheme (Ꜧ, $\underline{\text{Ḉ}}$, $\underline{\text{Đ}}$). Let $\text{Ă} = \{\langle \boldsymbol{j}, \boldsymbol{\mu}_{\text{Ă}}(\boldsymbol{j}), \boldsymbol{\nu}_{\text{Ă}}(\boldsymbol{j})\rangle, \boldsymbol{j} \in \text{Ꜧ}\}$ be an IF set in Ꜧ, where $\boldsymbol{\mu}_{\text{Ă}}(\boldsymbol{j})\text{: Ꜧ} \to [\mathbf{0}, \mathbf{1}]$ and $\boldsymbol{\nu}_{\text{Ă}}(\boldsymbol{j})\text{: Ꜧ} \to [\mathbf{0}, \mathbf{1}]$ are membership degree and non-membership grade of an example $\boldsymbol{j}$ to set Ă, in such a way that $\mathbf{0} \leq \boldsymbol{\mu}_{\text{Ă}}(\boldsymbol{j}) + \boldsymbol{\nu}_{\text{Ă}}(\boldsymbol{j}) \leq \mathbf{1}$. Consequently, hesitancy degree is described as the degree of uncertainty, where an example $\boldsymbol{j}$ has its place to set Ă is $\boldsymbol{\pi}_{\text{Ă}}(\boldsymbol{j}) = \mathbf{1} - \boldsymbol{\mu}_{\text{Ă}}(\boldsymbol{j}) - \boldsymbol{\nu}_{\text{Ă}}(\boldsymbol{j})$, such that $\mathbf{0} \leq \boldsymbol{\pi}_{\text{Ă}}(\boldsymbol{j}) \leq \mathbf{1}, \forall \boldsymbol{j} \in \text{Ꜧ}$.

An IF assisted binary relation ꝥ on Ꜧ can be explained by:

$$\text{ꝥ} = \{< (j,k), \mu_{\text{ꝥ}}(j,k), \vartheta_{\text{ꝥ}}(j,k) > \big| (j,k) \in \text{Ꜧ} \times \text{Ꜧ}\}$$

where, $\mu_{\text{ꝥ}}$: $\text{Ꜧ} \times \text{Ꜧ} \to [0,\ 1]$ and $\vartheta_{\text{ꝥ}}$: $\text{Ꜧ} \times \text{Ꜧ} \to [0,\ 1]$ hold $\mu_{\text{ꝥ}}(j,k) + \vartheta_{\text{ꝥ}}(j,k) \leq 1$ for any $(j,k) \in \text{Ꜧ} \times \text{Ꜧ}$. $\mu_{\text{ꝥ}}(j,k)\ and\ \vartheta_{\text{ꝥ}}(j,k)$ represents the essential similarity, and the important diversity degree of j to k respectively.

For two distinct essential IF relations ꝥ and Ḟ, where ꝥ is finer rather than Ḟ, implied by ꝥ ≤ Ḟ, iff $\mu_{\text{ꝥ}}(j,k)\ \leq \mu_{\text{Ḟ}}(j,k)$ and $\vartheta_{\text{ꝥ}}(j,k) \geq \mu_{\text{Ḟ}}(j,k)$ for any $(j,k) \in \text{Ꜧ} \times \text{Ꜧ}$. Let us consider $\text{Ꜧ} = \{j_1, j_2, \dots \dots .., j_n\}$. Hence a matrix $\text{ꝥ} = \left[\left(\mu_{\text{ꝥ}}(j_n, j_m), \vartheta_{\text{ꝥ}}(j_n, j_m)\right)\right]_{k \times k}$ can be applied to indicate an IF relation, where the $(n,m)^{th}$ value depicts the IF number between $j_n\ and\ j_m$.

Now, for the above mentioned IF set, an Intuitionistic fuzzy entropy [65] can be given as follows:

$$E_{\text{z}}(\text{ Ă}) = \sum_{q=1}^{k} \frac{1 - max\,\{\mu_{\text{Ă}}(j_q), \vartheta_{\text{Ă}}(j_q)\}}{1 - min\,\{\mu_{\text{Ă}}(j_q), \vartheta_{\text{Ă}}(j_q)\}}$$

3.1. IF Rough Set

Definition 1: Let Ꜧ be the set of whole data points and an IF relation be ꝥ. For any IF set $\text{Ă} \in IF$, a tuple namely lower and upper approximations of ẞ associated to above ꝥ on the basis of above discussed IF entropy can be calculated as mentioned follows:

$$\underline{ꝓ}(ẞ) = \left\{\frac{(\mu_{\underline{ꝓ}(ẞ)}(j), \vartheta_{\underline{ꝓ}(ẞ)}(j))}{j} | j \in Ꜧ\right\}$$

$$\overline{ꝓ}(ẞ) = \left\{\frac{(\mu_{\overline{ꝓ}(ẞ)}(j), \vartheta_{\overline{ꝓ}(ẞ)}(j))}{x} | j \in Ꜧ\right\}$$

where,

$$\mu_{\underline{ꝓ}(ẞ)}(j) = inf_{y \in U} \max\left(E_{\mathrm{z}}(\text{ Ă})\right)$$

$$\vartheta_{\underline{ꝓ}(ẞ)}(j) = sup_{y \in U} \min\left(E_{\mathrm{z}}(\text{ Ă})\right)$$

$$\mu_{\overline{ꝓ}(ẞ)}(j) = sup_{y \in U} \min\left(E_{\mathrm{z}}(\text{ Ă})\right)$$

$$\vartheta_{\overline{ꝓ}(ẞ)}(j) = inf_{y \in U} \max\left(E_{\mathrm{z}}(\text{ Ă})\right)$$

$\left(\mu_{\underline{ꝓ}(ẞ)}(j), \vartheta_{\underline{ꝓ}(ẞ)}(j)\right)$ tuple indicate IF numbers of a particular example j to the computed lower approximation assessment, whilst $(\mu_{\overline{ꝓ}(ẞ)}(j), \vartheta_{\overline{ꝓ}(ẞ)}(j))$ tuple indicate the IF numbers of a given example j to the computed upper approximation assessment. Then, $\left(\mu_{\underline{ꝓ}(ẞ)}(j), \vartheta_{\underline{ꝓ}(ẞ)}(j)\right)$ is considered as a required IF rough set [38, 39, 41, 43, 55, 57, 58]. Based on this lower approximation, positive region is identified. This positive region is further used to eliminate redundant samples by using a threshold value. Moreover, dependency function is established by computing ratio of cardinality of positive region and cardinality of the information system. Based on the information from dependency function reduct is computed by discarding redundant and/or irrelevant features. Due to use of IF entropy, uncertainty and noise are also handled successfully.

3.2. Support Vector Machine

This work identified support vector machine (SVM) [9, 18] as the best performing learning algorithm to build the prediction models. SVM has been utilized as the highly dominating machine learning method to solve biological prediction problems. This package comprises distinct kernels and machine learning is executed on the basis of information available in these kernels, where every dot entry is transmuted into non-linearity associated kernel function. RBF is a widely discussed squared exponential kernel, where abundant and flexible functional space is provided rather than a polynomial and/or linear kernel space. This produces much better results as elaborated in the literature. Collection of intake features with a fixed vector length are prerequisite for training phase, consequently, daunting a tactic for summarizing the overall information related to proteins/peptides of a given length format. After completion of training step, learned models can be used to perform the prediction of unseen data points [42].

3.3. WEKA Classifiers

Current study incorporates diverse classifiers available in WEKA [17] to carry out training of models, which is further used to perform the prediction of unknown instances. The distinct classifiers are applied to investigate the performance measures namely SMO, MLP, J48, decision tree, Random Forest, IBK, and PART. Different parameters are tuned during the performance of multiple machine learning methods, and we record the results attained on the best parameters.

3.4. Performance Measures

The performance of individual classification process is measured mainly by using two different type of measures, which is elaborated as threshold-dependent and threshold independent. In the current paper, we apply both above mentioned measures to perform the evaluation of various models. These measures can be easily and effectively enumerated from the estimates of the confusion matrix, viz.: true positives (Ŧp), which is the count of essential suitably predicted inducing peculiar peptides, false negatives (ʃn), which is the count of important incorrectly predicted inducing peculiar peptides, true negatives (Ŧn), which is the number of interesting correctly non-inducing peculiar peptides, and false positives (ʃp), which is the interesting number of incorrectly predicted non-inducing peculiar peptides.

Sensitivity (Sen)**:** This calculates the effective percentage of correctly predicted inducing peculiar peptides, and is specified by:

$$Sen = \frac{\text{Ŧp}}{(\text{Ŧp} + ƒ\text{n})} \times 100$$

Specificity (Spec)**:** This includes the efficacious percentage of correctly predicted non-inducing peculiar peptides, and is produced by:

$$Spec = \frac{\text{Ŧn}}{(\text{Ŧn} + \text{ʃp})} \times 100$$

Accuracy (Acc)**:** The percentage of very much required correctly predicted inducing and non-inducing peculiar peptides, and stated as:

$$Acc = \frac{\text{Ŧp} + \text{Ŧn}}{\text{Ŧp} + \text{ʃn} + \text{Ŧn} + \text{ʃp}} \times 100$$

AUC: It is applied to identify the much required the area under the receiver operating characteristic curve (ROC), the more devoted its count towards 1, the better will be the built predictor.

MCC**:** Mathew's correlation coefficient is an important and the most promising parameters, which evaluated by using the below given equation:

$$MCC = \frac{\text{Ŧp} \times \text{Ŧn} - \text{ʃn} \times \text{ʃp}}{\sqrt{(\text{Ŧp} + \text{ʃp})(\text{Ŧp} + \text{ʃn})(\text{Ŧn} + \text{ʃp})(\text{Ŧn} + \text{ʃn})}}$$

This parameter is used to clarify whether the binary classifications is more effective or less effective. An MCC value is more devoted towards 1 is used to clarify that the predictor is best one.

4. Result and Discussion

In the current section, we conduct the experiment by using 5 benchmark datasets [44] as well as Jain et al. dataset [1]. Whole experimental work is conducted by using 10-fold cross validation and a percentage split of 70:30. Features are extracted on the basis of valuable information from iFeature web server. Eleven distinct types of attributes are populated from the peptide sequences from Jain et al. dataset. Firstly, SMOTE is incorporated over all the six datasets to generate relative balanced dataset to proceed further. Entire details related to characteristics of all the six datasets along with their details after the implementation of SMOTE are outlined in Table I. Next, redundant data points are discarded by our proposed entropy based method. Then, redundant and irrelevant features are effectively eliminated from all the six datasets. Now, we use seven broadly used learning algorithms on these reduced datasets. These algorithms are an extension of SVM as Sequential minimization optimization (SMO) with RBF kernel, rotation forest, Random forest, MLP, IBK, J48, and (4) PART. Among these seven learning methods, SMO has performed much better when compared to other six methods. Bireduct method was implemented in Jupyter Notebook on specific hardware platform consisted of Intel(R) Core(TM) i3-5005U CPU @ 2.00 GHz with 4.00 GB RAM.

Details of all the six datasets are highlighted in Table I after the application of our proposed bireduct and recent fuzzy rough bireduct approach. Overall accuracy along with standard deviation is recorded for widely used SMO and PART machine leaning technique is recorded in Table II for the reduced datasets as produced by both our proposed approach and recently presented existing technique. From the experimental results as reported in Table I and Table II, our proposed methodology is provided better results whether the proposed approach has produced relatively less number of instances and/or features. Furthermore, an application of the proposed methodology is applied to enhance the discrimination performance of inducing and non-inducing peculiar peptides. Firstly, fixed feature vector of length 953 features was generated from IL-13 dataset consisted of inducing and non-inducing peculiar peptides by using iFeature web server. Then, SMOTE is incorporated to produce balanced dataset. Next, redundant samples are avoided by using our proposed entropy based intuitionistic fuzzy assisted concept. Moreover, redundant and irrelevant attributes are repudiated by using our proposed feature selection method. During entire process noise is eliminated as we apply entropy aided methods. All the seven machine learning techniques are investigated over reduced IL-13 dataset. This extensive experimental results are given in Tables III-VI. The best results are obtained based on the data reduction provided by our proposed approach with specificity, sensitivity, accuracy, MCC, and AUC as 84.0%, 87.9%, 87.7%, 0.697, and 0.898 respectively, which is much better when compared to previously reported results. Receiver optimization characteristics curve (ROC curve) is one of the widely discussed visualization technique for better understanding the experimental results from the numerous learning algorithms. For our experimental results, Figures 2-3 are the ROC curves for four learning algorithms based on reduced datasets produced by the existing fuzzy set based approach and our proposed approach, which clearly depicts the superiority of our presented model.

Table I

Benchmark information systems characteristics and their reduction

Dataset	Instances	Attributes	Balanced Dataset		Reduction			
					Fuzzy set assisted bireduct		Entropy based Intuitionistic Fuzzy set assisted bireduct	
			Instances	Attributes	Instances	Attributes	Instances	Attributes
Arcene	200	10000	217	10000	214	223	211	165
Bank-marketing	4521	21	79844	21	55565	18	52223	14
Breast-cancer	699	9	916	9	728	7	698	5
Colon	62	32	80	32	73	23	68	17
Ionosphere	351	33	445	33	399	25	372	18
IL-13	3221	953	5349	953	4998	343	4788	256

Table II
Comparison of overall classification/learning accuracies along with standard deviation for reduced datasets produced by fuzzy set based bireduct and entropy based intuitionistic fuzzy assisted bireduct using 10-fold cross validation

Dataset	Classification/Learning algorithms	Fuzzy Set based Bireduct	Entropy based Intuitionistic fuzzy set assisted Bireduct
Arcene	SMO	91.28 ± 6.55	92.33 ± 4.42
	PART	84.44 ± 5.22	86.17 ± 3.99
Bank-marketing	SMO	82.33 ± 1.22	83.89 ± 0.55
	PART	87.88 ± 3.11	89.77 ± 1.29
Breast-cancer	SMO	91.28 ± 2.12	92.97 ± 1.11
	PART	92.54 ± 3.68	95.43 ± 2.23
Colon	SMO	83.24 ± 9.12	84.88 ± 6.19
	PART	84.12 ± 10.47	87.34 ± 9.87
Ionosphere	SMO	82.33 ± 5.78	83.87 ± 3.55
	PART	86.44 ± 4.34	89.28 ± 2.43
IL-13	SMO	81.2 ± 2.33	84.7 ± 1.11
	PART	79.8 ± 2.11	81.7 ± 1.23

Table III

Performance evaluation metrics/parameters of learning methods for the discrimination of reduced inducing and non-inducing peculiar peptides consisted dataset produced by fuzzy set based bireduct using percentage split of 80:20

Learning methods	Sensitivity	Specificity	Accuracy	MCC	AUC
SMO	82.6	81.9	82.6	0.565	0.821
Random Forest	81.2	80.9	78.7	0.511	0.822
IBK	78.3	77.1	77.1	0.433	0.721
Decision Tree	74.5	73.2	77.1	0.477	0.731
J48	75.2	74.3	78.3	0.421	0.729
PART	78.9	79.1	78.6	0.511	0.799
MLP	80.2	80.1	79.2	0.525	0.801

Table IV
Performance evaluation metrics/parameters of learning methods for the discrimination of inducing and non-inducing peculiar peptides consisted dataset produced by fuzzy set based bireduct using 10-fold cross validation

Learning methods	Sensitivity	Specificity	Accuracy	MCC	AUC
SMO	83.1	81.9	84.2	0.572	0.833
Random Forest	81.9	81.2	80.3	0.565	0.821
IBK	79.1	78.3	80.2	0.525	0.781
Decision Tree	75.2	73.7	75.9	0.521	0.773
J48	77.2	75.9	77.9	0.559	0.698
PART	78.3	79.5	79.8	0.578	0.778
MLP	79.2	78.3	77.8	0.591	0.745

Table V

Performance evaluation metrics/parameters of learning methods for the discrimination of reduced inducing and non-inducing peculiar peptides consisted dataset produced by Entropy assisted Intuitionistic fuzzy set based bireduct using percentage split of 80:20

Learning methods	Sensitivity	Specificity	Accuracy	MCC	AUC
SMO	87.8	85.7	84.8	0.611	0.891
Random Forest	82.1	83.4	82.3	0.589	0.799
IBK	81.2	80.9	81.6	0.599	0.772
Decision Tree	78.6	79.2	80.1	0.612	0.778
J48	79.5	80.1	81.1	0.575	0.777
PART	81.2	82.1	80.7	0.577	0.801
MLP	81.9	82.1	81.4	0.498	0.787

Table VI

Performance evaluation metrics/parameters of learning methods for the discrimination of reduced inducing and non-inducing peculiar peptides consisted dataset produced by Entropy assisted Intuitionistic fuzzy set based bireduct using 10-fold cross validation

Learning methods	Sensitivity	Specificity	Accuracy	MCC	AUC
SMO	87.9	84.0	87.7	0.697	0.898
Random Forest	84.8	84.7	84.5	0.622	0.878
IBK	83.2	82.4	82.2	0.611	0.811
Decision Tree	80.1	79.8	80.1	0.609	0.833
J48	79.9	80.1	82.2	0.589	0.801
PART	82.2	82.3	81.7	0.599	0.803
MLP	82.1	83.1	81.6	0.598	0.823

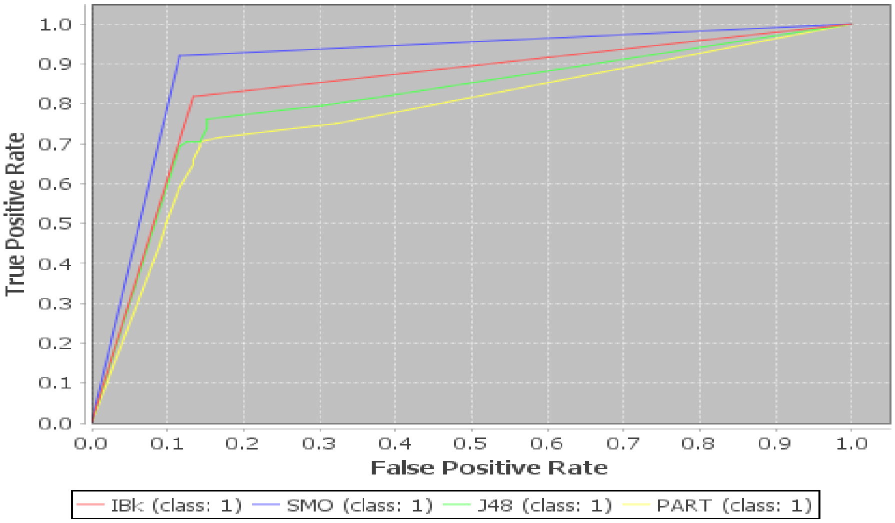

Fig 2. AUC of four machine learning algorithms for reduced IL-13 dataset by fuzzy rough set aided approach by using on 10-fold cross validation.

5. Conclusion

Different advancement in internet based technology and laboratory technology are the main cause of large volume of data production. Data is enlarging in the form of both size and dimensions. These data are composed of several issues such as redundancy in instances and dimensionality, irrelevant features, and noise. In this paper, SMOTE produced balanced data in the initial phase. Further, an entropy abetted intuitionistic fuzzy rough model was introduced. Furthermore, lower approximation was used to define positive region aided sample elimination method to remove redundant samples in the data. Moreover, the same positive region was employed to present degree of dependency concept to discard redundant and/or irrelevant attributes. Thereafter, numerous machine learning algorithms were analyzed on reduced data. From the observation of recorded results, it can be easily identified that our proposed methodology is performing better when compared to fuzzy based techniques. Moreover, we applied this methodology for an intriguing discrimination of induced peculiar and non-induced peculiar peptides, which is utilized for Covid-19 treatment.

In the future, we intend to improve this entropy based approach by information gain abetted approach. Moreover, probabilistic variable precision aided intuitionistic fuzzy approach can be presented to establish a bireduct method. Next, type-2 fuzzy set theory can be incorporated to establish an improved bireduct technique.

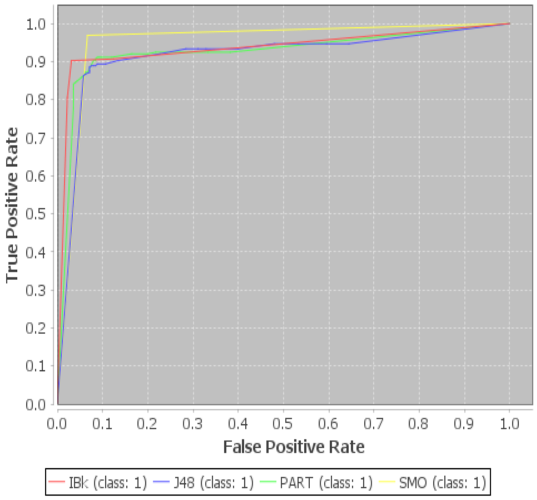

Fig 3. AUC of four machine learning algorithms for reduced IL-13 dataset by entropy aided intuitionistic fuzzy rough set based approach by using 10-fold cross validation

Acknowledgement: The authors are very much thankful to the Dr. Anoop Kumar Tiwari, Central University of Haryana, for his precious supports and comments to complete this work.

Reference

[1] Jain, Shipra, et al. "IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides." *Computers in Biology and Medicine* 143 (2022): 105297.

[2] Arora, Pooja, et al. "iIL13Pred: improved prediction of IL-13 inducing peptides using popular machine learning classifiers." *BMC bioinformatics* 24.1 (2023): 1-22.

[3] Minty, A., et al. "lnterleukin-13 is a new human lymphokine regulating inflammatory and immune responses." *Nature* 362.6417 (1993): 248-250.

[4] Q. A. Al-Radaideh, and G. Y. Al-Qudah, "Application of rough set-based feature selection for Arabic sentiment analysis," Cognitive Computation, vol. 9, pp. 436-445, 2017.

[5] K. T. Atanassov, On intuitionistic fuzzy sets theory: Springer, 2012.

[6] K. T. Atanassov, and K. T. Atanassov, Intuitionistic fuzzy sets: Springer, 1999.

[7] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," Computational Statistics & Data Analysis, vol. 143, pp. 106839, 2020.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[9] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert systems with applications, vol. 38, no. 7, pp. 9014-9022, 2011.

[10] J. Chen, J. Mi, and Y. Lin, "A graph approach for fuzzy-rough feature selection,"

Fuzzy Sets and Systems, vol. 391, pp. 96-116, 2020.

[11] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, and K.-C. Chou, “iFeature: a python package and web server for features extraction and selection from protein and peptide sequences,” Bioinformatics, vol. 34, no. 14, pp. 2499-2502, 2018.

[12] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data,” IEEE Transactions on Neural Networks and Learning Systems, 2022.

[13] Wynn, Thomas A. "IL-13 effector functions." *Annual review of immunology* 21.1 (2003): 425-456.

[14] D. Dubois, and H. Prade, “Rough fuzzy sets and fuzzy rough sets,” International Journal of General System, vol. 17, no. 2-3, pp. 191-209, 1990.

[15] D. Dubois, and H. Prade, “Putting rough sets and fuzzy sets together,” Intelligent decision support: Handbook of applications and advances of the rough sets theory, pp. 203-232, 1992.

[16] Junttila, Ilkka S. "Tuning the cytokine responses: an update on interleukin (IL)-4 and IL-13 receptor complexes." *Frontiers in immunology* 9 (2018): 888.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10-18, 2009.

[18] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18-28, 1998.

[19] B. Huang, Y.-L. Zhuang, H.-X. Li, and D.-K. Wei, “A dominance intuitionistic fuzzy-rough set approach and its applications,” Applied Mathematical Modelling, vol. 37, no. 12-13, pp. 7128-7141, 2013.

[20] P. Jain, and T. Som, “Multigranular rough set model based on robust intuitionistic fuzzy covering with application to feature selection,” International Journal of Approximate Reasoning, vol. 156, pp. 16-37, 2023.

[21] P. Jain, A. K. Tiwari, and T. Som, “A fitting model based intuitionistic fuzzy rough feature selection,” Engineering Applications of Artificial Intelligence, vol. 89, pp. 103421, 2020.

[22] P. Jain, A. K. Tiwari, and T. Som, “Enhanced prediction of anti-tubercular peptides from sequence information using divergence measure-based intuitionistic fuzzy-rough feature selection,” Soft Computing, vol. 25, pp. 3065-3086, 2021.

[23] P. Jain, A. K. Tiwari, and T. Som, “An intuitionistic fuzzy bireduct model and its application to cancer treatment,” Computers & Industrial Engineering, vol. 168, pp. 108124, 2022.

[24] R. Jensen, “Rough set-based feature selection: A review,” Rough computing: theories, technologies and applications, pp. 70-107, 2008.

[25] R. Jensen, and Q. Shen, “Fuzzy–rough attribute reduction with application to web categorization,” Fuzzy sets and systems, vol. 141, no. 3, pp. 469-485, 2004.

[26] R. Jensen, and Q. Shen, “Fuzzy-rough sets assisted attribute selection,” IEEE Transactions on fuzzy systems, vol. 15, no. 1, pp. 73-89, 2007.

[27] R. Jensen, and Q. Shen, “New approaches to fuzzy-rough feature selection,” IEEE Transactions on fuzzy systems, vol. 17, no. 4, pp. 824-838, 2008.

[28] W. Ji, Y. Pang, X. Jia, Z. Wang, F. Hou, B. Song, M. Liu, and R. Wang, “Fuzzy rough sets and fuzzy rough neural networks for feature selection: A review,” Wiley Interdisciplinary Reviews: Data Mining and

Knowledge Discovery, vol. 11, no. 3, pp. e1402, 2021.

[29] X. Ji, J. Peng, P. Zhao, and S. Yao, “Extended Rough Sets Model Based on Fuzzy Granular Ball and Its Attribute Reduction,” Information Sciences, pp. 119071, 2023.

[30] U. M. Khaire, and R. Dhanalakshmi, “Stability of feature selection algorithm: A review,” Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 4, pp. 1060-1073, 2022.

[31] X.-A. Ma, and Y. Yao, “Min-max attribute-object bireducts: On unifying models of reducts in rough set theory,” Information Sciences, vol. 501, pp. 68-83, 2019.

[32] N. Mac Parthaláin, R. Jensen, and R. Diao, “Fuzzy-rough set bireducts for data reduction,” IEEE Transactions on Fuzzy Systems, vol. 28, no. 8, pp. 1840-1850, 2019.

[33] M. M. Mafarja, and S. Mirjalili, “Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection,” Soft Computing, vol. 23, no. 15, pp. 6249-6265, 2019.

[34] Cho, Young Ae, and Jeongseon Kim. "Association of IL4, IL13, and IL4R polymorphisms with gastrointestinal cancer risk: a meta-analysis." *Journal of Epidemiology* 27.5 (2017): 215-220.

[35] Z. Pawlak, “Rough sets,” International journal of computer & information sciences, vol. 11, pp. 341-356, 1982.

[36] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, and W. Ziarko, “Rough sets,” Communications of the ACM, vol. 38, no. 11, pp. 88-95, 1995.

[37] Z. Qiu, and H. Zhao, “A fuzzy rough set approach to hierarchical feature selection based on Hausdorff distance,” Applied Intelligence, vol. 52, no. 10, pp. 11089-11102, 2022.

[38] S. Shreevastava, S. Singh, A. Tiwari, and T. Som, “Different classes ratio and Laplace summation operator based intuitionistic fuzzy rough attribute selection,” Iranian Journal of Fuzzy Systems, vol. 18, no. 6, pp. 67-82, 2021.

[39] S. Singh, S. Shreevastava, T. Som, and P. Jain, “Intuitionistic fuzzy quantifier and its application in feature selection,” International Journal of Fuzzy Systems, vol. 21, pp. 441-453, 2019.

[40] L. Sun, S. Si, W. Ding, X. Wang, and J. Xu, “TFSFB: Two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data,” Information Fusion, vol. 95, pp. 91-108, 2023.

[41] A. Tan, W.-Z. Wu, Y. Qian, J. Liang, J. Chen, and J. Li, “Intuitionistic fuzzy rough set-based granular structures and attribute subset selection,” IEEE Transactions on Fuzzy Systems, vol. 27, no. 3, pp. 527-539, 2018.

[42] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, “Comprehensive review on twin support vector machines,” Annals of Operations Research, pp. 1-46, 2022.

[43] A. K. Tiwari, S. Shreevastava, T. Som, and K. K. Shukla, “Tolerance-based intuitionistic fuzzy-rough set approach for attribute reduction,” Expert Systems with Applications, vol. 101, pp. 205-212, 2018.

[44] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007).

[45] C. Wang, Y. Huang, W. Ding, and Z. Cao, “Attribute reduction with fuzzy rough self-information measures,” Information Sciences, vol. 549, pp. 68-86, 2021.

[46] C. Wang, Y. Huang, M. Shao, Q. Hu, and D. Chen, “Feature selection based on neighborhood self-information,” IEEE Transactions on Cybernetics, vol. 50, no. 9, pp. 4031-4042, 2019.

[47] C. Wang, Y. Qian, W. Ding, and X. Fan, “Feature selection with fuzzy-rough minimum classification error criterion,”

IEEE Transactions on Fuzzy Systems, vol. 30, no. 8, pp. 2930-2942, 2021.

[48] X. Yang, H. Chen, T. Li, and C. Luo, “A noise-aware fuzzy rough set approach for feature selection,” Knowledge-Based Systems, vol. 250, pp. 109092, 2022.

[49] X. Yang, H. Chen, T. Li, P. Zhang, and C. Luo, “Student-t kernelized fuzzy rough set model with fuzzy divergence for feature selection,” Information Sciences, vol. 610, pp. 52-72, 2022.

[50] T. Yin, H. Chen, Z. Yuan, T. Li, and K. Liu, “Noise-resistant multilabel fuzzy neighborhood rough sets for feature subset selection,” Information Sciences, vol. 621, pp. 200-226, 2023.

[51] L. Zadeh, “Fuzzy sets,” Inform Control, vol. 8, pp. 338-353, 1965.

[52] J. Zhan, and B. Sun, “Covering-based intuitionistic fuzzy rough sets and applications in multi-attribute decision-making,” Artificial Intelligence Review, vol. 53, no. 1, pp. 671-701, 2020.

[53] A. Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, “SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors,” Information Sciences, vol. 595, pp. 70-88, 2022.

[54] X. Zhang, C. Mei, J. Li, Y. Yang, and T. Qian, “Instance and Feature Selection Using Fuzzy Rough Sets: A Bi-Selection Approach for Data Reduction,” 2022.

[55] X. Zhang, B. Zhou, and P. Li, “A general frame for intuitionistic fuzzy rough sets,” Information Sciences, vol. 216, pp. 34-49, 2012.

[56] Y. Zhang, D.-w. Gong, X.-z. Gao, T. Tian, and X.-y. Sun, “Binary differential evolution with self-learning for multi-objective feature selection,” Information Sciences, vol. 507, pp. 67-85, 2020.

[57] L. Zhou, and W.-Z. Wu, “On generalized intuitionistic fuzzy rough approximation operators,” Information Sciences, vol. 178, no. 11, pp. 2448-2465, 2008.

[58] L. Zhou, W.-Z. Wu, and W.-X. Zhang, “On intuitionistic fuzzy rough sets and their topological structures,” International Journal of General Systems, vol. 38, no. 6, pp. 589-616, 2009.

[59] H.-J. Zimmermann, Fuzzy set theory—and its applications: Springer Science & Business Media, 2011.

[60] Verbiest, Nele, Chris Cornelis, and Francisco Herrera. "FRPS: a fuzzy rough prototype selection method." *Pattern Recognition* 46.10 (2013): 2770-2782.

[61] Jensen, Richard, and Chris Cornelis. "Fuzzy-rough instance selection." *International Conference on Fuzzy Systems*. IEEE, 2010.

[62] Zhang, Xiao, et al. "A fuzzy rough set-based feature selection method using representative instances." *Knowledge-Based Systems* 151 (2018): 216-229.

[63] Zhang, Xiao, et al. "Active incremental feature selection using a fuzzy-rough-set-based information entropy." *IEEE Transactions on Fuzzy Systems* 28.5 (2019): 901-915.

[64] Anaraki, Javad Rahimipour, et al. "A Fuzzy-Rough Feature Selection Based on Binary Shuffled Frog Leaping Algorithm." *International Journal of Computer and Information Engineering* 12.9 (2018): 722-729.

[65] Szmidt, Eulalia, and Janusz Kacprzyk. "Entropy for intuitionistic fuzzy sets." *Fuzzy sets and systems* 118.3 (2001): 467-477.