# A Fuzzy Aggregation based Ensemble Framework for Accurate and Stable Feature Selection

Zixiao Shen, Xin Chen, Jonathan M. Garibaldi
*Intelligent Modelling and Analysis Group, School of Computer Science*
*Lab for Uncertainty in Data and Decision Making (LUCID)*
*University of Nottingham*, Nottingham, NG8 1BB, United Kingdom
{Zixiao.Shen, Xin.Chen, Jon.Garibaldi}@nottingham.ac.uk

***Abstract*—A novel ensemble feature selection (FS) framework using fuzzy aggregation is proposed in this paper. It consists of three main steps, including distribution generation of feature importance, distribution ensemble using fuzzy aggregation and defuzzification for feature ranking. Based on four state-of-the-art FS methods (named as base selectors in our algorithm) selected from different method categories, different fuzzy aggregation operators were implemented to achieve ensemble learning for decision making. A training data repository that consists of eight datasets was used for parameter tuning of the proposed framework. The proposed framework using drastic sum aggregation achieved the best performance and was subsequently evaluated on eight independent testing datasets. Remarkably, the proposed method achieved overall the best classification accuracy and the highest stability in comparison with the four base FS methods. It also outperformed our previously proposed score based ensemble method [1].**

## I. Introduction

Due to the rapid development and wide application of information technology, an increasing amount of data with high complexity is generated. This leads to curse of dimensionality which makes it much difficult to understand the underlying relationship accurately [2]. Feature selection (FS) methods are commonly chosen as a critical step to help design accurate and robust machine learning models in this situation. FS methods aim to select a feature subset for efficiently representing the input data, reducing the effects from noise or irrelevant features, and finally achieve precise decision making [3].

Comparing with the other FS approaches, such as wrapper and embedded methods, filter FS methods are independent of any learning algorithms. They include feature ranking as a principle or auxiliary selection mechanism on account of its simplicity, scalability and empirical success [3]. Through measuring the feature quality based on the filter method, feature ranking is widely applied as the essential step of FS. Based on the feature ranking results, the top ranked features are selected, while the number of features to select is specified by the user or analytically determined [4]. Comparatively, these approaches are highly computationally efficient, hence adopted in this research.

Rather than focus on increasing the performance of the downstream decision making methods, more and more attention is paid to enhance the stability of FS methods. Improvement on stability of FS methods not only help us select the relevant features with higher confidence, but also reduce the time consumption of acquiring new data in many practical applications, such as cancer detection, gene selection and etc. Application domain experts can also obtain the quantified evidence for the results' reliability from the analysis of FS methods' stability [1]. Hence, one of the objectives of this research is to develop a performance measurement to comprehensively evaluate FS methods on both accuracy and stability.

Another major challenge in FS area lies in the inconsistency of results using different FS methods [5]. From the previous research [1], the implementation of several filter FS methods on the same datasets may lead to various feature ranking sequences. This illustrated the high discrepancies and uncertainties of results by applying different FS methods. In the literature, fuzzy theory acts as an unified framework to model the vagueness, imprecision and uncertainty information within data. Through transforming the values into more than one label using various membership functions, vagueness and uncertainty within data are modelled and further exploited to enable reasoning [6].

Ensemble learning using filter FS methods is able to reduce the variations and increase the reliability and stability of FS results [5]. In our previous research [1], a novel weighted combination method for FS using score based approach was introduced. The method produced comparable classification accuracies with significantly higher stability than the base selectors, when variations and size reduction were introduced to the data. To improve and investigate the performance on different datasets, the previous research was extended with the following new contributions.

- Four base selectors are carefully selected from different categories within filter FS methods, which helps in generating decisions from various aspects.
- Normalized fuzzy sets are utilized to represent the distributions of the features' importance. These fuzzy sets are then combined using fuzzy aggregation operator to achieve ensemble learning. It has shown to improve the stability of feature ranking.
- The proposed method was thoroughly evaluated on eight testing datasets with different characteristics. It outperformed the base FS methods and the score based ensemble method [1], in terms of classification accuracy and stability.

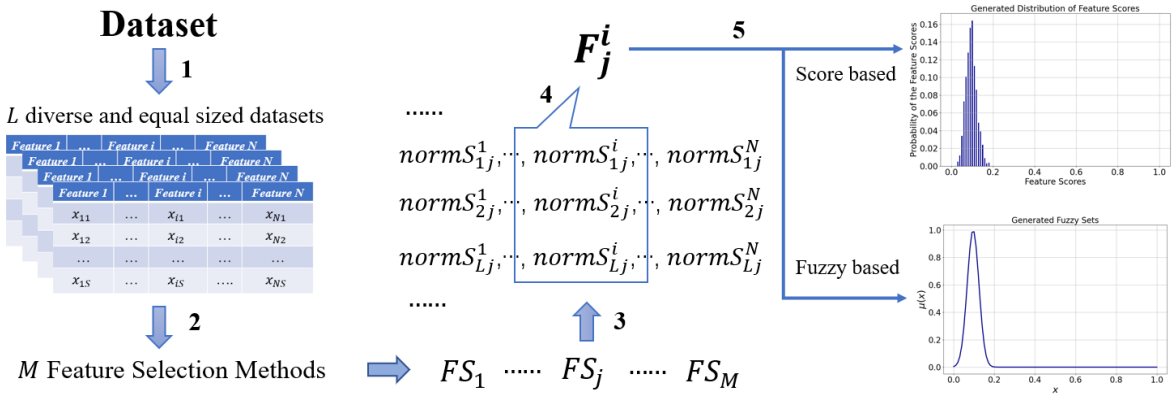



Fig. 1. Process of the distribution generation of feature importance.

## II. Methodology

A similar ensemble framework is adopted from our previous research [1]. Three main steps are followed in the framework, including: (1) distribution generation of feature importance; (2) distribution ensemble using fuzzy aggregation; (3) defuzzification for feature ranking. A score based distribution generation and weighted combination methods were used in our previous research [1]. In contrast, alternative fuzzy-based distribution generation and aggregation method are proposed in this paper, as illustrated in Fig. 2.

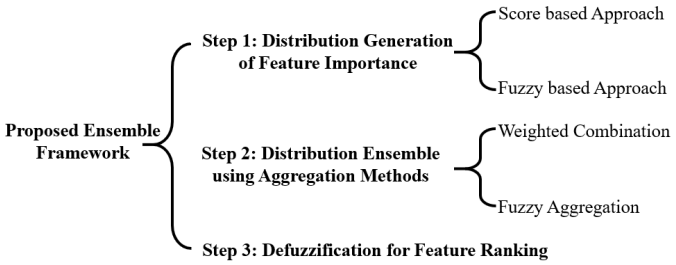



Fig. 2. Overview of the proposed ensemble framework.

### A. Distribution Generation of Feature Importance

To increase the generalizability of FS methods, bootstrap aggregation process is firstly applied to construct a number (denoted as $L$) of data subsets with $N$ features. $M$ different FS methods (denoted as $FS_j, j \in [1, ..., M]$) are then applied on these subsets to generate the feature scores for each subset per method. Distributions are constructed using $L$ bootstrap subsets to represent the importance of each feature. An overview of the distribution generation of feature importance is illustrated in Fig. 1. Detailed procedures are shown below.

*1) **Bootstrap Aggregation Process**:* Randomly select the samples and generate $L$ diverse and equal sized data subsets with replacement (denoted as $Subsetl, (l \in [1, ..., L])$).

*2) **FS Method Deployment**:* Apply FS method $FS_j$ on $Subset\ l$ to get the feature score (denoted as $S^i_{j,l}$) of the $i$th feature ($i \in [1, ..., N]$).

*3) **Feature Score Normalization**:* Normalize the feature scores of each FS method to the range of [0, 1] using min-max normalization [7] expressed in Equation (1).

$$normS^i_{j,l} = \frac{S^i_{j,l} - min\{S_j\}}{max\{S_j\} - min\{S_j\}} \tag{1}$$

where $min\{S_j\}$ and $max\{S_j\}$ represent the min and max values of the feature scores generated by $FS_j$ for all features.

*4) **Feature Discretization**:* Each of the feature space is divided into 101 equal sized interval scores (set as $U$).

$$U = \{0, 0.01, 0.02, ..., 0.98, 0.99, 1\} \tag{2}$$

Previously normalized feature scores are discretized and mapped into the corresponding element of set $U$ (denoted as $dS^i_{j,l}$), then combined into a list $F^i_j = \{dS^i_{j,l} | l \in [1, ..., L]\}$.

*5) **Feature Importance Representation**:* Probability density function $PDF^i_j(x)$ (Equation (3)) represents the feature scores which indicate the importance of the corresponding feature. The feature scores with higher PDF values are more representative for the features' importance.

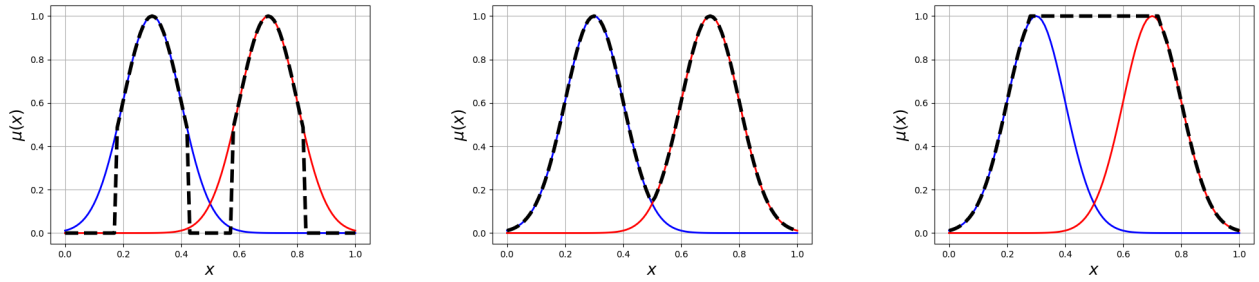
$$PDF^i_j(x) = \begin{cases} freq(x, F^i_j)/L & x \in F^i_j \\ 0 & x \notin F^i_j \end{cases} \tag{3}$$

On account of the bootstrap process, the FS calculation of different subsets is regarded as independent and identically distributed (i.i.d) experiments. Based on Bernoulli's law of large numbers [8], for any positive number $\epsilon$, we have

$$\lim_{L \to \infty} P\{|\frac{freq(x, F^i_j)}{L} - \mu| < \epsilon\} = 1 \tag{4}$$

Equation (4) indicates that when $L$ becomes large enough, $freq(x, F^i_j)/L$ can be used to represent the membership function values. The generation procedures are shown as below.

1) Given FS method $FS_j$ and the $i$th feature, calculate the mean value $mean^i_j$, standard deviation $\delta^i_j$ and height $height^i_j$ from the probability density function $PDF^i_j(x)$.
2) Constitute a normalized type-1 fuzzy set to represent the distribution of importance on the $i$th feature and $FS_j$, expressed in Equation (5).




(a) Yager class with $\omega = 1000$ (b) Yager class with $\omega = 10$ (b) Drastic sum

Fig. 3. Demonstration of fuzzy aggregation. Red and blue lines show the original fuzzy sets. Black dashed line indicates the fuzzy sets after aggregation.

$$Dist_j^i = \{(x, \mu_j^i(x)) | x \in X\} \tag{5}$$

Based on Bernoulli's law of large numbers, distributions of membership functions are described as gaussian shaped. The membership function $\mu_j^i$ is constructed using Equation (6).

$$\mu_j^i(x) = \frac{1}{\sqrt{2\pi}\delta_j^i} exp(-\frac{(x - mean_j^i)^2}{2(\delta_j^i)^2}) \tag{6}$$

### B. Distribution Ensemble using Fuzzy Aggregation

Fuzzy aggregation acts as the linear extensions of Boolean connectives within the scale between 0 and 1 [10]. In this paper, different fuzzy aggregation methods are implemented to combine the fuzzy sets generated in Section II-A. In the literature, various fuzzy aggregation operators are proposed, such as T-norm and S-norms. T-norm operators normally produce the intersection among fuzzy sets and eliminate other information outside the region, which lead to the loss of information during the aggregation process. On the other hand, S-norms are the generalized form of fuzzy union by integrating different fuzzy sets together, which retain most parts of information and are chosen in this research.

There are different kinds of S-norms, such as drastic sum, Dombi class, Yager class and etc. A parameterized family of S-norms Yager class is chosen on account that it covers different situations within the value range. Given $a$ and $b$ representing the membership function values of different fuzzy sets, Yager class is represented in Equation (7).

$$s_\omega(a, b) = min[1, (a^\omega + b^\omega)^{1/\omega}], \quad \omega \in (0, \infty) \tag{7}$$

In the extreme case when $\omega$ becomes 0, Yager class turns to be drastic sum s-norm $s_{ds}(a, b)$.

$$s_{ds}(a, b) = \begin{cases} a & if\ b = 0 \\ b & if\ a = 0 \\ 1 & otherwise \end{cases} \tag{8}$$

The combined fuzzy sets after the aggregation process using Yager class with different parameters ($\omega = 1000$, 10) and drastic sum S-norms are visualized in Fig. 3.

### C. Defuzzification for Feature Ranking

*1)* ***Defuzzification****:* After combining different fuzzy sets, a single value is computed using defuzzification processes. Centroid defuzzifier, or also called center of gravity defuzzifier is applied in the process [11].

*2)* ***Feature Ranking****:* The feature ranking sequence is obtained from the highest to lowest values using the final defuzzified feature scores. Afterwards, the ranking sequence is utilized as the guidance for the subsequent decision making process.

## III. Method Evaluation

In this section, datasets, experimental design, evaluation metrics and evaluation results are described.

### A. Datasets

*1)* ***Training Data Repository****:* Eight datasets from UCI machine learning repository [9] with different levels of data sparsity are chosen to tune the parameters of the proposed method. General information is shown in Table I.

TABLE I
General Description of Training Data Repository

| ID | Datasets | $\#C$ | $\#F$ | $\#S$ | $\#S/\#C/\#F$ |
|---|---|---|---|---|---|
| 1 | Banknote | 2 | 4 | 1372 | 171.5 |
| 2 | Mammographic | 2 | 5 | 830 | 83 |
| 3 | PIMA | 2 | 8 | 768 | 48 |
| 4 | Statlog Heart | 2 | 13 | 270 | 10.4 |
| 5 | Seeds | 3 | 7 | 210 | 10 |
| 6 | Sports Articles | 2 | 59 | 1000 | 8.5 |
| 7 | Parkinsons | 2 | 22 | 195 | 4.4 |
| 8 | Spectfheart | 2 | 44 | 267 | 3.03 |

$\#C$, $\#F$ and $\#S$ represent the number of classes, features and samples respectively. In the following tables, the datasets are represented using the ID numbers in the first column accordingly.

*2)* ***Testing Data Repository****:* Another 8 datasets with different data sparsity are chosen for independent evaluation, as shown in Table II. Those datasets are used for final performance analysis and comparison.

TABLE II
GENERAL DESCRIPTION OF TESTING DATA REPOSITORY

| ID | Datasets | $\#C$ | $\#F$ | $\#S$ | $\#S/(\#C \times \#F)$ |
|---|---|---|---|---|---|
| 9 | Yeast | 10 | 8 | 1484 | 18.55 |
| 10 | CMSC | 2 | 18 | 540 | 15 |
| 11 | WDBC | 2 | 30 | 569 | 9.48 |
| 12 | Appendicitis | 2 | 7 | 106 | 7.57 |
| 13 | BCC | 2 | 9 | 116 | 6.44 |
| 14 | Glass | 6 | 9 | 214 | 4.0 |
| 15 | Breast Tissue | 6 | 9 | 106 | 2.0 |
| 16 | Musk | 2 | 166 | 476 | 1.43 |

### B. Selection of FS Methods

*1) **Base Selectors**:* In our experiments, four representative algorithms (achieved superior performance in the literature) are selected from different filter FS categories [12], which are implemented as the base selectors. General description of the chosen base selectors is shown in Table III.

TABLE III
GENERAL DESCRIPTION OF BASE SELECTORS

| No. | Alias | Name | Category |
|---|---|---|---|
| 1 | CFS | Correlation based FS [13] | Statistical-based FS |
| 2 | ReliefF | ReliefF FS [14] | Similarity-based FS |
| 3 | MIFS | Mutual Information based FS [15] | Information-based FS |
| 4 | IFS | Infinite based FS [16] | Graph-based FS |

*2) **The Proposed Ensemble Methods**:* Through incorporating different aggregation techniques, S-norms in Yager class (denoted as $F\_Snorm$) and the drastic sum fuzzy aggregation method (denoted as $F\_DS$) are used in our experiments. The performance of the proposed fuzzy based approaches are also compared with the score based method (denoted as $S\_OW$). $S\_OW$ method utilized the score based approach to generate the distributions and One Minus Standard Deviation Weights (OW) method for weighted combination, which is the best method reported in our previous research [1].

### C. Evaluation Metrics

Two metrics are used to evaluate the performance of FS methods on the aspects of accuracy and stability.

*1) **Evaluation on Accuracy**:* FS methods are ultimately used to improve the down-stream decision making performance (e.g. classification accuracy in our experiments) with simpler model and reduced number of features. After the pre-processing process using FS methods, the feature ranking sequence is obtained from the most to the least significant. Through gradually eliminating the unimportant features, predictive performance of FS methods is measured when combining with different decision making techniques. Detailed procedures are shown below.

1) Divide the data into the training set and testing set using $K$-fold cross validation;
2) Implement the given FS method on the training set to rank the features;
3) Remove the least important feature one at a time. Use the retained features to train a model on the training set and make prediction on the testing set;
4) For each reduced sized feature set, calculate the mean accuracy of all folds.

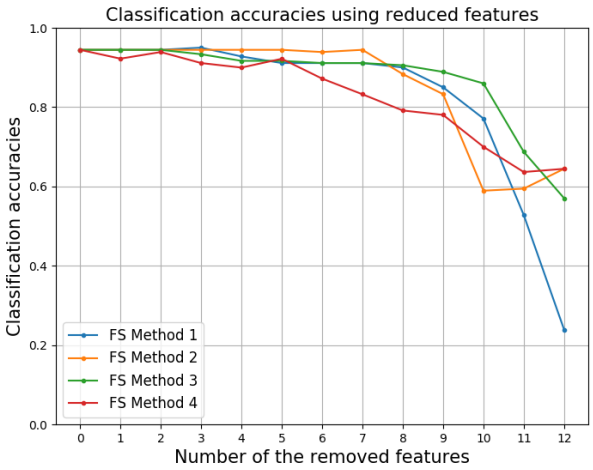



Fig. 4. Visualization of accuracies using removed features.

After implementing those procedures, the accuracies using reduced feature sets and different FS methods can be visualized, as the example shown in Fig. 4.

For quantitative method comparison, a single value is normally summarized from these plots in Fig. 4, such as mean, maximum or weighted sum [17]. Here, we proposed a new evaluation metric in Equation (9), which is similar to the weighted sum [17] in the calculation process and principle. However, this metric takes advantages of normalizing the performance score to the range of [0, 1], which is suitable for performance comparison among different datasets.

$$Weighted\ Acc. = \frac{\sum Acc. * \%RemovedFeatures}{\sum \%RemovedFeatures} \quad (9)$$

where $\%RemovedFeatures$ represents the proportion of removed features. $Acc.$ indicates the classification accuracies based on the retained features. Based on Equation (9), the accuracies using higher proportion of removed features have larger weights than others.

*2) **Evaluation on Stability**:* To evaluate the stability of a given algorithm, recent research mainly focuses on the stability indices and introduces the metrics based on Hamming distance, correlation coefficients, consistency, information theory and etc [18]. A consistent feature ranking is more important than a fixed feature score in the context of FS. Hence rather than using Pearson's correlation to evaluate the feature score [1], Spearman's Rank Correlation Coefficient (SRCC) is implemented in this study to measure the difference between the ranking sequences generated from different folds in the cross-validation experiments.

$$ASC = \frac{\sum_{x=1}^{K}\sum_{y=x+1}^{K} SRCC(r^x, r^y)}{K(K-1)/2} \quad (10)$$

$SRCC(r_x, r_y)$ represents the SRCC between feature ranking sequences $r^x$ and $r^y$. $K$ stands for the number of folds. Equation (10) measures the average SRCC value for all pairwise comparisons of different folds.

## IV. Experiments & Results

### A. Parameter Tuning using Training Data Repository

In order to investigate the behavior of the proposed method when the parameter changes, the proposed method is firstly applied to the training data repository using 10-fold cross validation ($K = 10$) for parameter tuning. Based on the sorted features, Random Forest is selected to model the classification process and calculate the accuracies on account of its bagging character and robustness to outliers and non-linear data. Performance of the FS methods is reported based on both accuracy and stability.

For the aims of tuning and selecting an optimal fuzzy aggregation approach, performance within the proposed ensemble FS framework using different fuzzy aggregation methods was compared. S-norms of Yager class with different parameters and drastic sum were utilized as the aggregation approaches. Within Yager class S-norms, the parameter $\omega$ was set as 1000, 10, 0.1 and 0.001 respectively. To compare the performance in the extreme case when $\omega = 0$, drastic sum is also applied as a supplementary.

*1)* ***Parameter Tuning on Accuracy****:* Based on Section III-C1, comparison using accuracy evaluation metric is shown in Table IV.

TABLE IV
Performance on Accuracy using Fuzzy based Approach

| Data | $F\_Snorm$ with different $\omega$ 1000 | 10 | 0.1 | 0.001 | $F\_DS$ |
|---|---|---|---|---|---|
| 1 | **.864** (1.5) | .666 (4) | .673 (3) | **.864** (1.5) | **.864** |
| 2 | .769 (3) | .746 (4) | .772 (2) | **.827 (1)** | **.827** |
| 3 | .717 (3) | .706 (4) | **.744** (1) | .728 (2) | .728 |
| 4 | .705 (4) | .743 (1) | **.757** (1) | .730 (3) | .730 |
| 5 | .776 (3) | .756 (4) | .786 (2) | **.798** (1) | **.798** |
| 6 | .735 (4) | **.767** (1.5) | **.767** (1.5) | .761 (3) | .761 |
| 7 | .790 (3) | **.811** (1) | .793 (4) | .798 (2) | .798 |
| 8 | **.791** (1.5) | .788 (3) | .785 (4) | **.791** (1.5) | **.791** |
| AVG | .768 (2.9) | .748 (2.8) | .760 (2.3) | .787 (1.9) | .787 |

In Table IV, numbers in bold indicate the best performance for the given dataset. AVG means the average performance value among all the datasets in the table. The numbers in brackets represent the ranking index for $F\_Snorm$ using different parameters. Lower value indicates a higher rank and a better FS performance. In the case that more than one method obtained the same results, joint ranks were scored using the average value of the rank orders. For instance, ranking index 1 stands for the FS method achieved the best performance. When two methods achieved the same best performance, both of them were then scored 1.5.

It is seen from Table IV that $F\_Snorm$ achieved better performance with smaller parameter value $\omega$. When $\omega$ became close to 0 ($\omega = 0.001$ in this experiment), $F\_Snorm$ obtained the same optimal result as the drastic sum method $F\_DS$.

*2)* ***Parameter Tuning on Stability****:* Based on Section III-C2, comparison using stability evaluation metric is shown in Table V.

TABLE V
Performance on Stability using Fuzzy based Approach

| Data | $F\_Snorm$ with different $\omega$ 1000 | 10 | 0.1 | 0.001 | $F\_DS$ |
|---|---|---|---|---|---|
| 1 | **1.00** (2) | .960 (4) | **1.00** (2) | **1.00** (2) | **1.00** |
| 2 | .467 (4) | .958 (2) | .924 (3) | **1.00** (1) | **1.00** |
| 3 | .733 (4) | .961 (3) | **1.00** (1.5) | **1.00** (1.5) | **1.00** |
| 4 | .811 (4) | .842 (3) | .969 (2) | **1.00** (1) | **1.00** |
| 5 | .810 (4) | .914 (2) | .841 (3) | **1.00** (1) | **1.00** |
| 6 | .513 (4) | .869 (3) | **.985** (1) | .943 (2) | .943 |
| 7 | .696 (4) | .738 (3) | .808 (2) | **1.00** (1) | **1.00** |
| 8 | **1.00** (1.5) | .784 (4) | .932 (3) | **1.00** (1.5) | **1.00** |
| AVG | .754 (3.4) | .878 (3) | .932 (2.2) | **.993** (1.4) | **.993** |

In Table V, the stability performance of $F\_Snorm$ method increased significantly with the decrease of parameter $\omega$ values in Yager class S-norms. When $\omega = 0.001$, $F\_Snorm$ achieved the best performance in 7 out of 8 datasets, which is the same as the drastic sum method $F\_DS$.

In summary, the fuzzy based approach using $F\_Snorm$ with $\omega$ close to 0 or the drastic sum method $F\_DS$ achieved the best performance on both accuracy and stability. As drastic sum is a simpler expression comparing with Yager class S-norms, $F\_DS$ is then chosen for further performance analysis and comparison using the test datasets.

### B. Performance Analysis using Testing Data Repository

In this section, the performance of the fuzzy based approach is evaluated and analyzed on the testing data repository. To evaluate the effects of bootstrap aggregation process on FS results, performance of the proposed method is compared with the base selectors which utilized the same bootstrap aggregation process. Base selectors after bootstrap aggregation represent the FS methods which utilize bootstrap aggregation to generate the feature distributions and then defuzzify without any combination or aggregation procedures. In addition, the performance of the proposed fuzzy based approach is also compared with the score based approach in [1]. The approach using One Minus Standard Deviation Weights combination method ($S\_OW$) has achieved the best performance hence been chosen here.

*1)* ***Performance Analysis on Accuracy****:* Based on Section III-C1, comparison on accuracy is shown in Table VI. In the columns of base selectors after bootstrap aggregation, the operators (+/-) indicate that the bootstrap aggregation process has increased (+), decrease (-) or same (no mark) performance compared to the base selectors without bootstrap. The numbers in the brackets indicate the joint ranking index using the proposed method $S\_OW$ (before the slash) or $F\_DS$ (after the slash) comparing with the four base selectors after bootstrap

TABLE VI
PERFORMANCE ANALYSIS ON ACCURACY

| Data | Base Selectors with Bootstrap Aggregation | | | | Proposed Methods | |
|---|---|---|---|---|---|---|
| | CFS | ReliefF | MIFS | IFS | $S\_OW$ | $F\_DS$ |
| 9 | .439 (1.5/1.5) | .351 (5/5) | .439 (1.5/1.5) | .409 (3/3) | .399 (4/) | .391 (/4) |
| 10 | .915 (3/3) | .915 (3/3) | .915 (3/3) | .915 (3/3) | .915 (3/) | .915 (/3) |
| 11 | .920 (4.5/4.5) | $.920^{-}$ (4.5/4.5) | .926 (1.5/1) | .922 (3/3) | .926 (1.5/) | .925 (/2) |
| 12 | $.823^{-}$ (2/3) | $.836^{-}$ (1/1) | $.819^{+}$ (3/4) | $.806^{+}$ (5/5) | .817 (4/) | .832 (/2) |
| 13 | .307 (5/5) | $.467^{-}$ (2/3) | $.521^{-}$ (1/1) | $.404^{-}$ (3/4) | .319 (4/) | .519 (/2) |
| 14 | .349 (2/3) | .256 (5/5) | $.383^{+}$ (1/1) | $.305^{-}$ (4/4) | .309 (3/) | .364 (/2) |
| 15 | .267 (5/5) | $.303^{-}$ (1/2) | $.270^{+}$ (4/4) | $.297^{+}$ (3/3) | .301 (2/) | .340 (/1) |
| 16 | $.453^{-}$ (4/4) | $.440^{-}$ (5/5) | $.481^{-}$ (2.5/3) | $.485^{-}$ (1/2) | .481 (2.5/) | .528 (/1) |
| AVG | $.559^{-}$ (3.4/3.6) | $.561^{-}$ (3.3/3.5) | $.594^{+}$ (2.2/2.3) | .568 (3.1/3.4) | .559 (3/) | .602 (/2.1) |

aggregation. AVG stands for the average performance value for all datasets in terms of accuracy and ranking index.

In Table VI, our proposed methods $F\_DS$ produced overall better performance than $S\_OW$ and the four base selectors. It is also observed that the bootstrap aggregation for individual base selector does not lead to a higher accuracy in most cases for different datasets and FS methods. This indicates the improved accuracy mainly results from the proposed fuzzy aggregation method rather than the bootstrap process.

*2) **Performance Analysis on Stability**:* Based on Section III-C2, performance comparison on stability is shown in Table VII.

TABLE VII
PERFORMANCE ANALYSIS ON STABILITY

| Data | Base Selectors with Bootstrap Aggregation | | | | Proposed Methods | |
|---|---|---|---|---|---|---|
| | CFS | ReliefF | MIFS | IFS | $S\_OW$ | $F\_DS$ |
| 9 | 1.00 (2/2.5) | .995 (4.5/5) | $1.00^{+}$ (2/2.5) | 1.00 (2/2.5) | .995 (4.5/) | 1.00 (/2.5) |
| 10 | 1.00 (1/1.5) | $.854^{+}$ (3/4) | $.856^{+}$ (2/3) | $-.078^{+}$ (5/5) | .672 (4/) | 1.00 (/1.5) |
| 11 | $.949^{+}$ (5/5) | $1.00^{+}$ (1/1.5) | $.989^{+}$ (3/4) | $.995^{+}$ (2/3) | .988 (4/) | 1.00 (/1.5) |
| 12 | $.967^{+}$ (1/2) | $.919^{+}$ (3/4) | $.941^{+}$ (2/3) | $.917^{+}$ (4/5) | .815 (5/) | 1.00 (/1) |
| 13 | $.915^{+}$ (3/4) | $.965^{+}$ (2/3) | $.756^{+}$ (5/5) | $.993^{+}$ (1/2) | .905 (4/) | 1.00 (/1) |
| 14 | $.972^{+}$ (4/4) | $.980^{+}$ (1/2) | $.890^{+}$ (5/5) | .975 (3/3) | .976 (2/) | 1.00 (/1) |
| 15 | .667 (5/5) | $.980^{-}$ (1/2) | $.957^{+}$ (2/3) | $.944^{+}$ (3/4) | .922 (4/) | 1.00 (/1) |
| 16 | $.935^{+}$ (2/3) | $.887^{+}$ (4/4) | $.886^{+}$ (5/5) | .985 (1/2) | .919 (3/) | 1.00 (/1) |
| Avg. | $.925^{+}$ (2.9/3.4) | $.948^{+}$ (2.4/3.2) | $.909^{+}$ (3.3/3.8) | $.841^{+}$ (2.6/3.3) | .899 (3.8/) | 1.00 (/1.3) |

Comparing the performance of base selectors before and after bootstrap aggregation, the bagging process leads to increased stability in all the datasets. However, our proposed fuzzy based approach $F\_DS$ further outperformed the base selectors with bootstrap aggregation and $S\_OW$ in all datasets for stability measure.

In summary, the proposed fuzzy based approach is able to achieve the overall best classification accuracy and the highest stability compared to the base selectors, which also outperformed our previously proposed score based method [1].

## V. DISCUSSION & CONCLUSION

In this paper, an ensemble FS framework using fuzzy aggregation method has been proposed. By generating various distributions using the bootstrap approach, normalized fuzzy sets are utilized to represent the features' importance. Fuzzy aggregation approaches S-norms such as Yager class and drastic sum are implemented to aggregate the distributions from different FS methods. Through combining four state-of-the-art filter FS methods from different categories, our method produces a final score for feature ranking.

The proposed method was firstly evaluated on a training data repository for the purpose of parameter tuning for the proposed fuzzy aggregation method. It was subsequently evaluated on an independent testing data repository for performance comparison in terms of classification accuracy and stability. Our proposed method with drastic sum s-norms fuzzy aggregation has produced the best average classification accuracy of all the testing datasets. Additionally, it has also achieved significantly higher stability in all testing datasets.

In future work, we will compare the proposed method to other state-of-the-art FS approaches, and use more evaluation metrics such as robustness when incomplete data and outliers are considered.

## REFERENCES


[1] Z. Shen, X. Chen, and J. M. Garibaldi, "A novel weighted combination method for feature selection using fuzzy sets," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019, pp. 1–6.

[2] Z. Shen, X. Chen, and J. Garibaldi, "Performance optimization of a fuzzy entropy based feature selection and classification framework," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 1361–1367.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[4] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1399–1414, 2003.

[5] F. Kamalov and F. Thabtah, "A feature selection method based on ranked vector scores of features for classification," *Annals of Data Science*, vol. 4, no. 4, pp. 483–502, 2017.

[6] R. Jensen, "Combining rough and fuzzy sets for feature selection," Ph.D. dissertation, Citeseer, 2005.

[7] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[8] G. Grimmett, G. R. Grimmett, D. Stirzaker *et al.*, *Probability and random processes*. Oxford university press, 2001.

[9] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[10] R. Mesiar and A. Kolesárová, "Aggregation functions in fuzzy set theory: History and some recent advances," in *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*. IEEE, 2018, pp. 94–97.

[11] L.-X. Wang, *A course in fuzzy systems and control*. Prentice-Hall, Inc., 1996.

[12] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2018.

[13] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[14] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[15] M. Zaffalon and M. Hutter, "Robust feature selection by mutual information distributions," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 577–584.

[16] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4202–4210.

[17] Z. Shen, X. Chen, and J. M. Garibaldi, "A novel meta learning framework for feature selection using data synthesis and fuzzy similarity," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

[18] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, 2019.