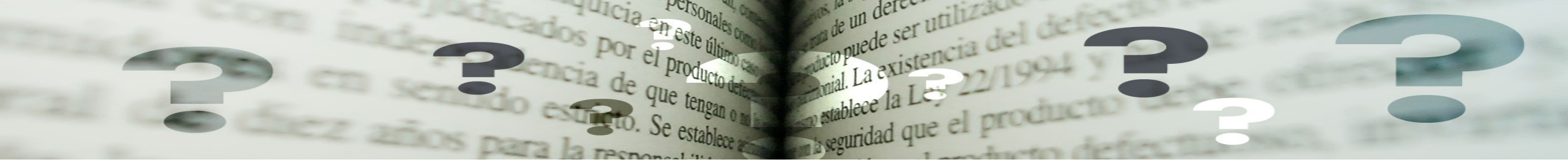


An Ensemble-based Feature Selection Methodology *for* Case-Based Learning

PhD. Dissertation Presentation

Maqbool Ali[1,2]

[1]Department of Computer Science and Engineering,
Kyung Hee University, South Korea
Email: maqbool.ali@oslab.khu.a.c.kr

[2]School of Engineering and ICT,
University of Tasmania, Australia
Email: maqbool.ali@utas.edu.ac

04th May, 2018

경희대학교
KYUNG HEE UNIVERSITY

Advisor:
(Prof. Sungyoung Lee)

UNIVERSITY OF TASMANIA

Advisor:
(Prof. Byeong Ho Kang)

Agenda

- Introduction
  - Background
  - Motivation
  - Problem statement
  - Research Taxonomy
- Related work
- Proposed methodology
  - Overview
  - Workflow
- Experiment & results
  - Dataset
  - Experimental setup
  - Results & discussion
- Conclusion
  - Contribution & Uniqueness
  - Future work
- Publications
- References

Background

- In medical education domain, Case-Based Learning (CBL) is known to be an *effective* learning approach for *medical students* at undergraduate level education as well as for *professional development* [1-3].
  - CBL is a shared learning approach in which small-groups of medical students are involved in discussion to *identify* and *solve* the patient's problem [1].
- In CBL practice,
  - the clinical case is a key component in learning activities, which includes *basic*, *social*, and *clinical* studies of the patient [1]. It provides a foundation to understand the situation of a disease.

Mr. X a 65 years old corporate sector person, came to medical expert with few complaints. On inquiring, he told that he is providing finance consultancy to the clients. He added that his office hrs are 8:30 am to 6:00 pm. As his job is related to office work. He has no physical activities. He used to drink regularly and like to eat fatty and oily food. According to him, he is used to tire quite early from last few weeks. He felt fatigue and breathlessness after a small walk even 100 meter. He has a problem of blurred vision along with weight loss. He told that he has never been in such problem before. He has not taken any medicine. His height is 183cm and weight is 89kg. His family has a history of hypertension and hyperglycemia. Expert worried about his health and alarmed him to be conscious about his health. For observing vital signs, expert suggested him to use wearable devices to register his blood pressure, glucose level, and heart rate. On Examination: Systolic Blood Pressure = 135.24 mmHg, Diastolic Blood Pressure = 89.33 mmHg, Glucose Level in fasting = 145.43 mg/dL, Glucose Level in random = 247.36 mg/dL, Heart Rate = 90.14 bpm

An example of a clinical case

Medical Students

Goal

- To interact with the patients
- To deal with a variety of cases during his/her clinical practical life
- Better learning can play an important role in actual practice

- Human can not
  - perform fast reasoning
  - accomplish complex computation decision

For better learning

- Structured knowledge can be:
  - ✓ Queried
  - ✓ Analyzed
  - ✓ Visualized

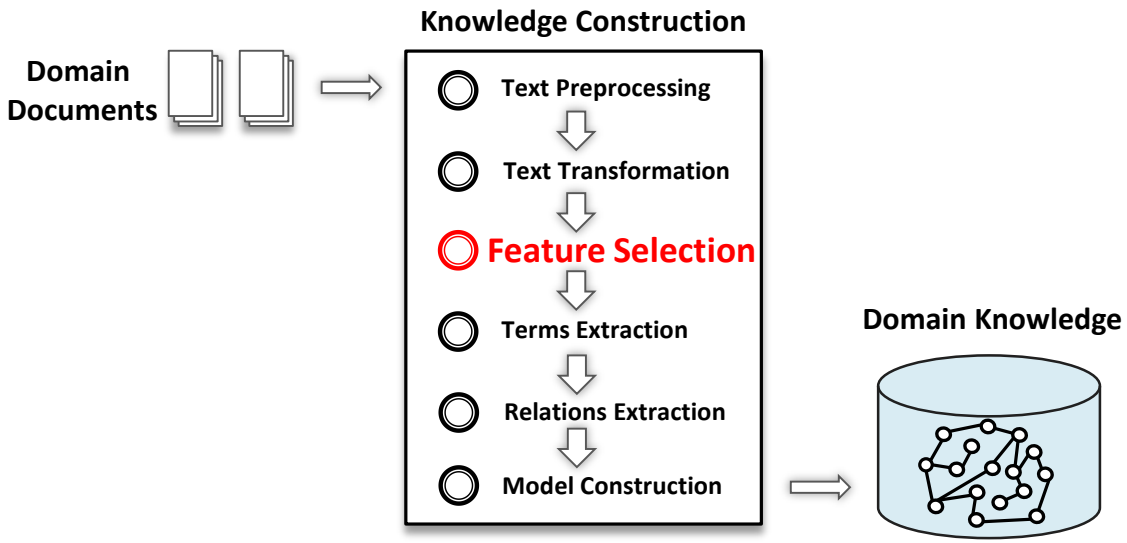
Domain Knowledge
(i.e. Structured Declarative Knowledge)

CBL

- **Better decision making**

Declarative knowledge is a type of knowledge, which tells us facts: what things are.
- "Blood disease is a symptom of diabetes"

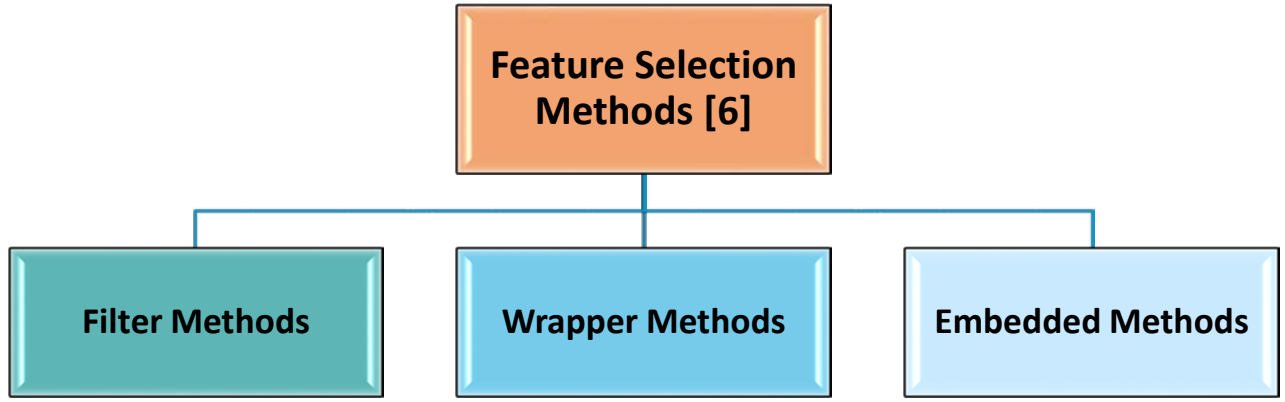
Background and Motivation

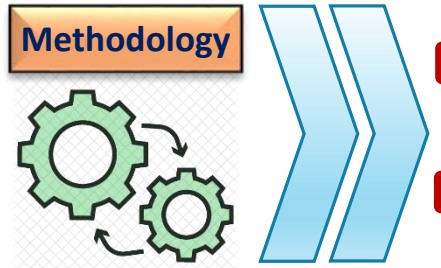


- Text Mining is the process of deriving high-quality information from an unstructured text [4]. It involves the application of techniques from *information retrieval*, *natural language processing*, *information extraction*, and *data mining*.
- For constructing domain knowledge,
  - Feature selection is an important and critical step in text mining [5].

R Large number of features selection methods available

R Each method has capabilities and limitations





O Comprehensive evaluation of feature set

O Ensemble feature selection

A Not dependent on the classification algorithm

A Better accuracy

Problem Statement

For an automated CBL, a reliable structured knowledge construction is a challenging task [7]. The key challenge in this regard is to *select* the relevant features for the following reasons:

- The irrelevant input features induces greater computational cost [6, 8].
- Finding an optimal cut-off value to select important features is problematic [9].

Goal

- Innovate students' learning by transforming the unstructured text into structured knowledge with the support of an efficient feature selection methodology.

Objectives

1. To *design* and *develop* an efficient *feature selection methodology* to filter out the irrelevant input features for structured knowledge construction process.
2. To *innovate* the *case-based learning* approach for better clinical proficiency.

Challenges

1. How to compute the ranks of features without any individual statistical biases of state-of-the-art feature ranking methods? [10] (e.g., information gain is biased towards choosing feature with large number of value. Similarly, chi square, symmetric uncertainty, and gain ratio are sensitive to sample size.
2. How to provide an empirical method to specify a minimum threshold value for retaining important features? [11]
3. How to design the *case-based learning* approach to make it interactive and effective? [12]

Research Taxonomy [13, 14]

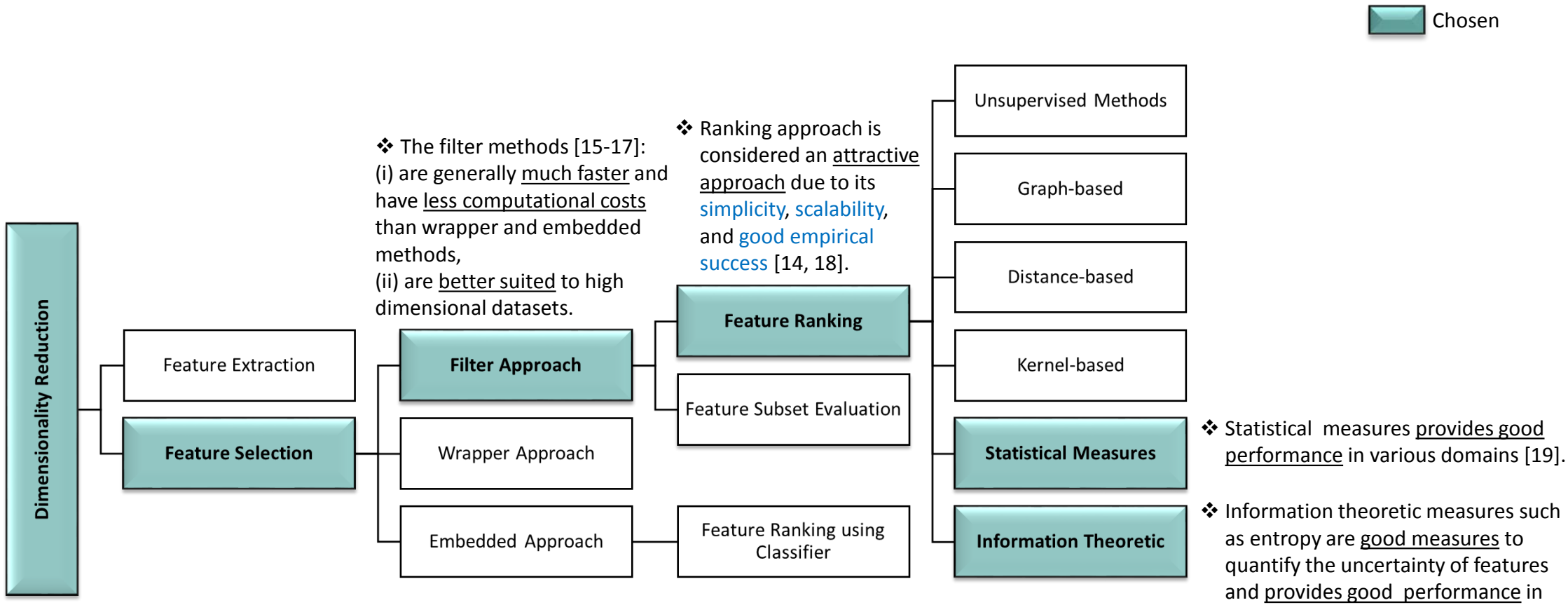


Figure: Dimensionality reduction and different categories of feature ranking methods.

Related Work

	Reference	Features	Limitations
Feature Selection	[20] Onan and Korukoğlu, A feature selection model based on genetic rank aggregation for text sentiment classification, 2017.	• Presented an ensemble approach for feature selection, which aggregates the several individual feature lists obtained by the different feature selection methods such as *Information gain*, *Gain ratio*, *Chi-squared, Pearson Correlation, ReliefF*. • Used Naïve Bayes and kNN classifiers	• Genetic algorithm (GA) was used for producing an aggregate ranked list, which is relatively *more expensive* technique than a weighted aggregate technique. • Experiments were primarily *performed a binary-class problem*. Hence, it is *not clear how* would the proposed method will deal with more complex datasets?
	[11] Osanaiye et al., Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing, 2016.	• Presented an ensemble-based multi-filter feature selection method that combines the output of *Information gain*, *Gain ratio*, *Chi-squared* and *ReliefF* to select important features.	• A *fixed threshold value* i.e. 1/3 of a feature set, was defined a priori irrespective of the characteristics of the dataset.
	[10] Sarkar et al., Robust feature selection technique using rank aggregation, 2014.	• Proposed a technique that aggregates the *Information gain*, *Chi-Square*, and *Symmetric Uncertainty* feature selection methods to develop an optimal solution.	• This technique is *not comprehensive* enough to provide a final subset of features. Hence, a domain expert would still *needed to make an educated guess* regarding the final subset.
	[13] Sadeghi and Beigy, A new ensemble method for feature ranking in text mining, 2013.	• Proposed a heterogeneous ensemble-based algorithm for feature ranking using *Information gain, Relief, and DRB-FS* features ranking methods. • Adopted *borda method* for features voting • Determined the threshold using genetic algorithm.	• This method *requires user to specify a* θ value. • Moreover, user is given an additional task of *defining the notion of relevancy* and *redundancy* of a feature. • The proposed wrapper-based method is *tightly coupled* with the performance evaluation of a *single classifier* i.e. SVM, hence *losing the generality* of the method.
Case-Based Learning	[21] University of Texas Medical Branch UTMB, Design a case (DAC), 2017.	• Provides facility to develop case(s) • Delivers virtual patient encounters to students on any health related topic • Support of anywhere accessible	• This approach *does not provide domain knowledge support* for CBL practice
	[22] The University of New Mexico, Extension for community healthcare outcomes (ECHO), 2016.	• Provides services for remote patient care • Conducts virtual clinics using multi-point videoconferencing	• *Lacks* of an *interactive case authoring* and its *formulation support* • *Lacks* of *domain knowledge support* for CBL practice
	[23] Chen et al., Applications of a time sequence mechanism in the simulation cases of a web-based medical problem-based learning system, 2009.	• Developed a web-based learning system that followed the development of the real-world clinical situation	• *Lacks* of *feedback support* • *Lacks* of *domain knowledge support* for CBL practice

Idea Diagram

Limitations

1 **[10, 11, 13, 20]** Existing methodologies

-- used relatively *more expensive* techniques to select the features **OR**

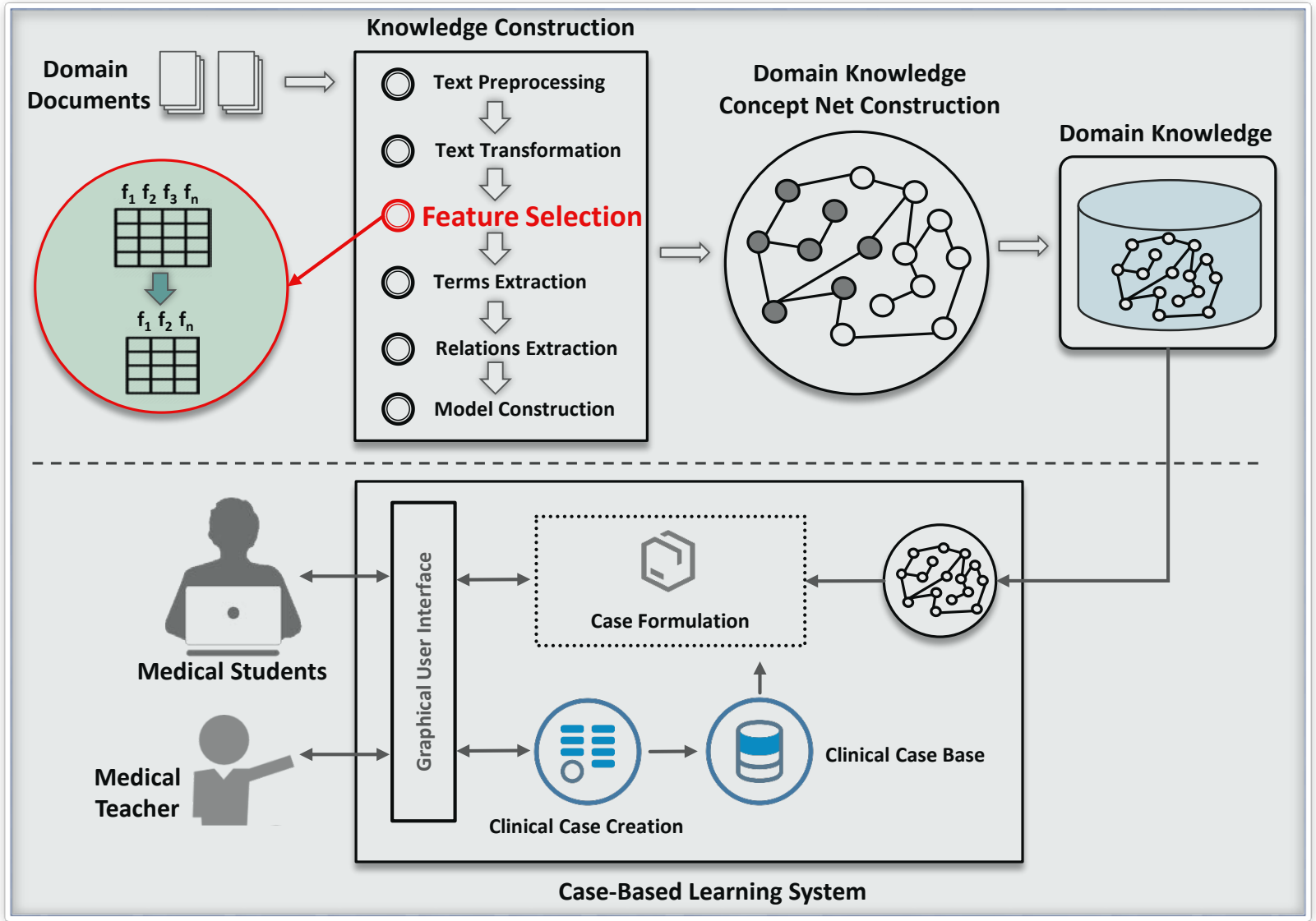
-- *required an educated guess* to specify a minimum threshold value for retaining important features

2 **[21, 22, 23]** Existing CBL approaches are designed:

-- *without* describing the procedures that how clinical cases are developed **OR**

-- *without* an interactive case authoring and a case formulation support **OR**

-- *without* domain knowledge support



Solutions

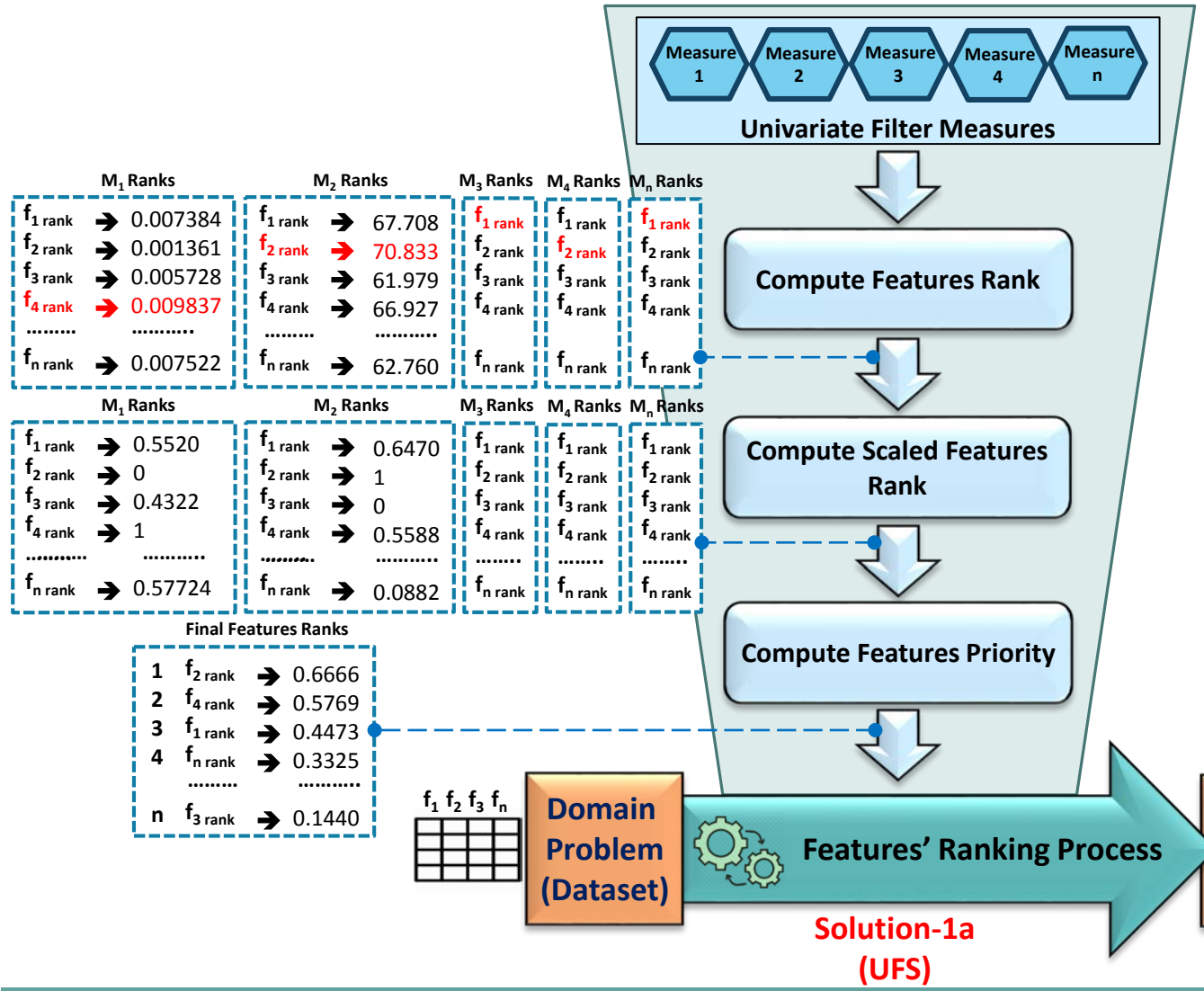
1a, 1b **1a.** Proposed a *flexible approach* for incorporating state-of-the-art univariate filter measures for feature ranking.

1b. Proposed an *efficient approach* for selecting a cut-off value for the threshold in order to select a subset of features.

2 Introduced an *effective* CBL approach using real-world clinical case creation and case formulation techniques

(Solution-1a & 1b)

Proposed *Univariate Ensemble-based Feature Selection* (uEFS) Methodology



Highlight of the idea

- Find the more appropriate features of a dataset
- Do the features' ranking process with a proposed *Unified Features Scoring* (UFS) algorithm
- Select the features using a proposed *Threshold Value Selection* (TVS) algorithm

Assumptions

1. Filter measures provide ranks in terms of numeric values
2. Our selected datasets with varied complexities represent a general case (Relatively balance dataset).

Detailed Workflow – (Solutions-1a & 1b)

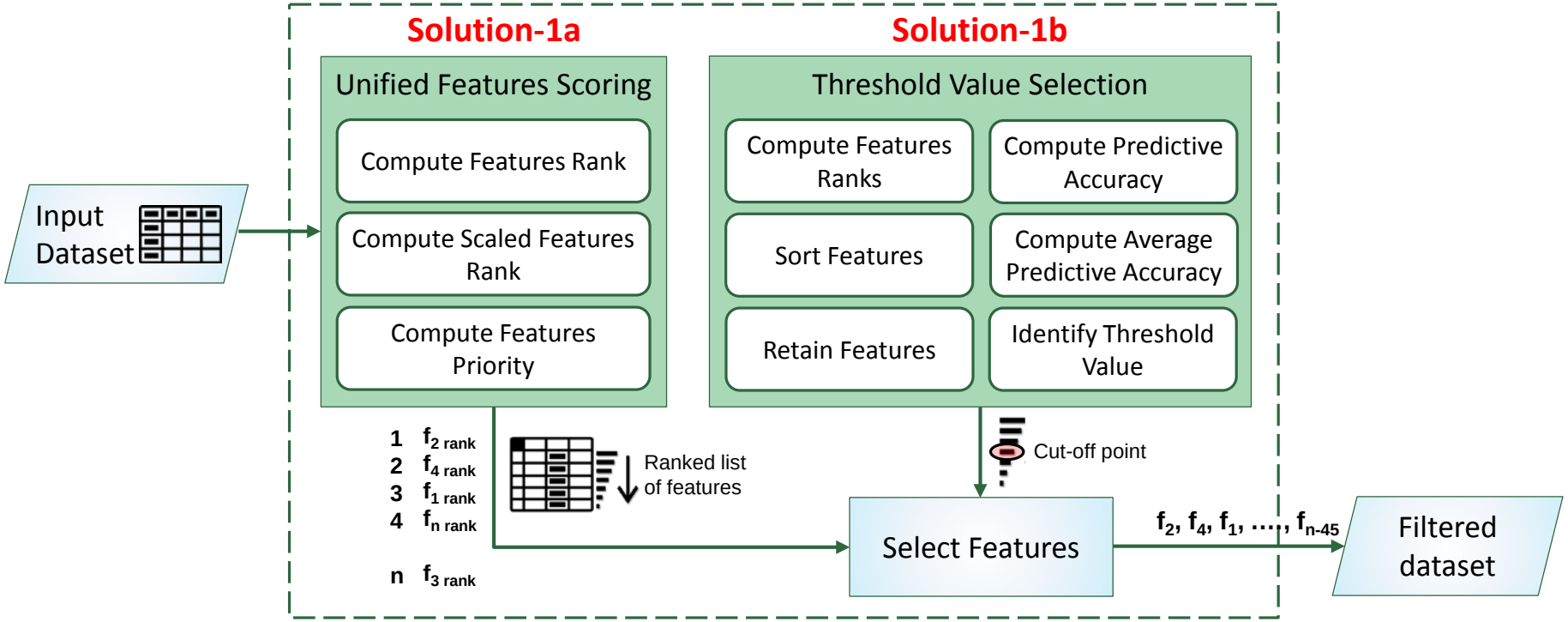


Figure: The detail workflow of the proposed *Univariate Ensemble-based Feature Selection* (uEFS) methodology

- In the proposed uEFS methodology, We contribute two components
  1. **Unified features scoring (UFS):** a comprehensive and flexible filter-based ensemble technique
  2. **Threshold value selection (TVS):** data characteristics guided threshold value selection

Different to existing approaches:

- **UFS** neutralizes the biasness of the state-of-the-art features ranking measures.
- **TVS** provides an empirical method of specifying a minimum threshold value to retain important features for decision making process.

Proposed *Unified Features Scoring* (UFS) Algorithm – (Solution-1a)

Input: Dataset

Output: Ranked Features Set

1. Compute the number of features
2. Compute the feature ranks using ***n*** number of univariate filter-based measures
3. Compute the scaled ranks for all computed ranks using the *Algorithm-2*
4. Compute the combined sum of all computed ranks
5. For each feature, add computed scaled ranks (from step-3)
6. Sort the ranks in ascending order
7. Compute the score, weight, and priority of each feature

Algorithm 2: Scaling the Computed Ranks (CR)

```
Data: CR: Input computed ranks (ranks)
Result: SR– Scaled Ranks
smallest ← ranks_0 ;
largest ← ranks_0 ;
for ∀ noOfAttrs ∈ CR do
    if rank_i > largest then
        largest ← rank_i
    else
        if rank_i < smallest then
            smallest ← rank_i
        end
    end
end
min ← smallest ;
max ← largest ;
SR[] ← (ranks − min)/(max − min) ;
return SR : scaled ranks
```

Reason for considering Filter-based method:

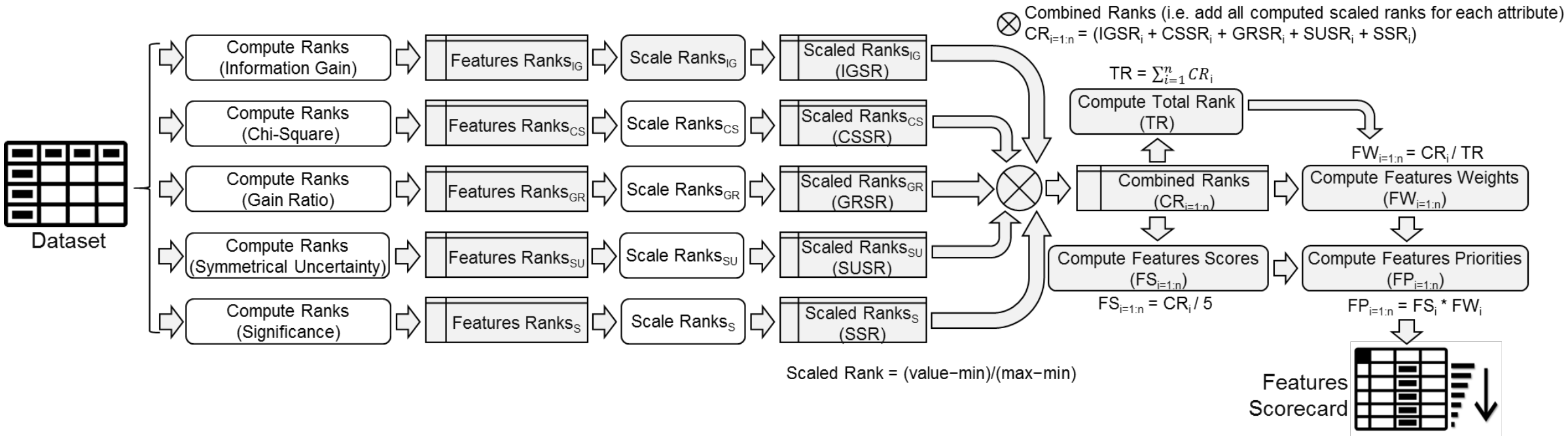
- **Why Filter Method? [6]**
  - ❖ This method performs simple and fast computation.
  - ❖ It does not depend on the classification algorithm.
  - ❖ Set of all features ➔ Selecting the best subset ➔ Learning Algorithm ➔ Performance
- **Why Univariate Filter Measures? [20]**
  - ❖ Have been widely utilized owing to their simplicity and relatively high performance.

Filter Algorithm

input: $D(F_0, F_1, ..., F_{n-1})$ // a training data set with N features
S_0 // a subset from which to start the search
δ // a stopping criterion
output: S_{best} // an optimal subset

```
01 begin
02     initialize: S_best = S_0;
03     γ_best = eval(S_0, D, M); // evaluate S_0 by an independent measure M
04     do begin
05         S = generate(D); // generate a subset for evaluation
06         γ = eval(S, D, M); // evaluate the current subset S by M
07         if (γ is better than γ_best)
08             γ_best = γ;
09             S_best = S;
10     end until (δ is reached);
11     return S_best;
12 end;
```

Proof of Concept *for* UFS algorithm – (Solution-1a)



Reason for considering following Univariate Measures for Features' Ranking Process:

- **Information Gain**: One of the popular measure used for feature selection, which informs features contribution in enhancing information about the target class [24].
- **CHI Squared**: Statistical measure that determines the association between feature and its class [24]
- **Gain Ratio**: One of disparity measures that enhances the Information Gain [24]
- **Symmetrical Uncertainty**: Performed well for highly imbalanced features set [25]
- **Significance**: Probabilistic measure that assess the feature's worth [26]

Selected Filter Measures for the UFS algorithm – (Solution-1a)

Information Gain (IG): [24]

- IG is an information theoretic measure, which is computed by following equation:
  - InformationGain(A) = Info(D) − InfoA(D) , where
- *InformationGain(A)* is the information gain of feature A, which is an independent attribute.
- *Info(D)* is the entropy of the entire dataset.
- InfoA(D) is the conditional entropy of feature A over D.

Gain Ratio (GR): [24]

- GR utilizes the split information value that is given as follows:
  - $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2 \frac{|D_j|}{|D|}$, where
- *SplitInfo* represents the structure of partitions.
- Finally, GR is defined as follows:
  - GainRatio(A) = InformationGain(A) / SplitInfo(A)

Symmetrical Uncertainty (SU): [25]

- SU is an information theoretic measure to assess the rating of constructed solutions. It is a expressed by the following equation:
  - $SU(A,B) = \frac{2 * IG(A|B)}{H(A) + H(B)}$, where
- *IG(A|B)* represents the information gain computed by independent feature *A* and class attribute *B*.
- *H(A)* and *H(B)* represent the entropies of feature *A* and *B*.

CHI Squared (CS): [24]

- CS helps to measure the independence of feature from its class. It is defined as:
  - $CHI(t,c_i) = \frac{N * (AD - BE)^2}{(A+E) * (B+D) * (A+B) * (E+D)}$, where

$$CHI_{max}(t) = \max_i(CHI(t,c_i))$$

- *A*, *B*, *E*, and *D* represent the frequencies of occurrence of both *t* and C_i, *t* without C_i, C_i without *t*, and neither C_i nor *t* respectively. While *N* represents the total number of features.

Significance (S): [26]

- The significance of an attribute A_i is denoted by $\sigma(A_i)$, which is computed by following equation:
  - $\sigma(A_i) = \frac{AE(A_i) + CE(A_i)}{2}$, where $AE(A_i)$ represents the cumulative effect of all possible attribute to class association of an attribute A_i, while $CE(A_i)$ represents the association between the attribute A_i and various class decisions.
  - $AE(A_i) = \left(1/k \sum_{r=1,2,\ldots,k} \vartheta_i{}^r\right) - 1.0$, where *k* represents the different values of attribute A_i. Similarly $CE+(A_i) = (1/m) * \left(\sum_{j=1,2,\ldots,m} A_i{}^j\right) - 1.0$, where *m* represents the number of classes, while $+(A_i)$ depicts the the class-to attribute association for the attribute A_i

Proposed *Threshold Value Selection* (TVS) Algorithm – (Solution-1b)

Input: Datasets

Output: Predictive accuracy graph to reveal the cut-off value

1. Consider ***n*** number of benchmark datasets having *varying complexities*
2. For each dataset:
   a) Compute the feature ranks using *Ranker Search* mechanism.
   b) Based on the computed ranks, sort all features in an ascending order
3. Partition each dataset into different chunks (filtered dataset) from 100% to 5% features retained
4. Feed each filtered dataset to ***m*** number of classifiers having *varying characteristics* (where ***m*** << ***n***)
5. Using *10-fold cross validation* approach, record *predictive accuracies* of these classifiers to each chunk of dataset partitioning
6. Compute *average predictive accuracy* of all classifiers as well as datasets against each chunk of dataset partitioning
7. Plot all computed average predictive accuracies against each chunk of dataset partitioning
8. Identify the cut-off value from plotted graph

Why Ranker Search mechanism?

- It is considered an optimal solution to score the features [27].

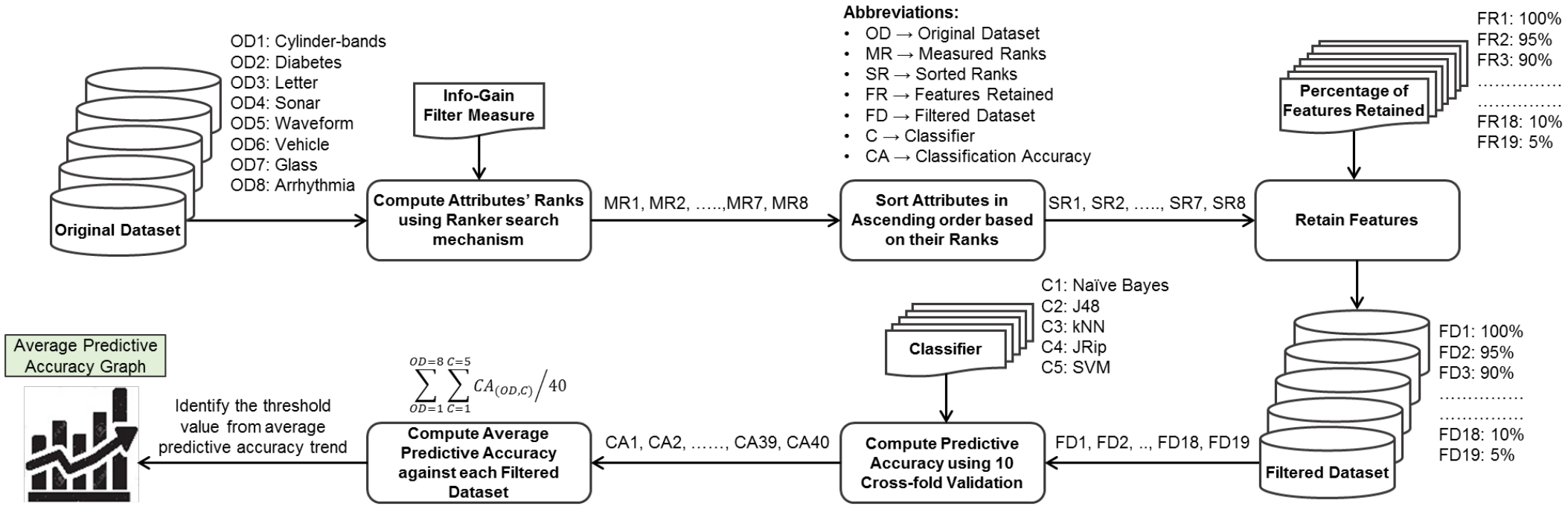
Why 10-fold Cross Validation?

- Most commonly used approach for model validation [28, 29].

Main intuitions of this algorithm are:

- To identify an appropriate chunk value that will provide reasonable predictive accuracy
- To specify those attributes which are deemed important for the domain construction
- To reduce the dataset

Proof of Concept *for* TVS algorithm – (Solution-1b)



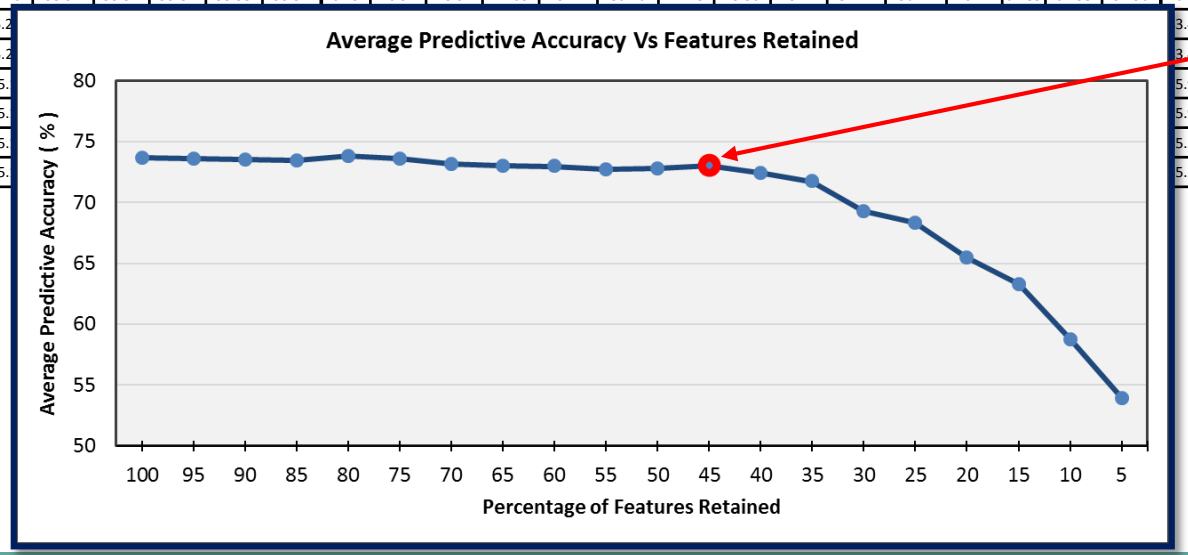
- Considered eight benchmark UCI datasets of *varying complexities* (no. of classes)
  1. Cylinder-bands (2)
  2. Diabetes (2)
  3. Letter (2)
  4. Sonar (2)
  5. Waveform (3)
  6. Vehicle (4)
  7. Glass (6)
  8. Arrhythmia (13)

- Considered five well-known classifiers having *varying characteristics* (classifier family/category)
  1. Naïve Bayes (Bayes)
  2. J48 (Trees)
  3. kNN (Lazy)
  4. JRiP (Rules)
  5. SVM (Functions)

Proof of Concept *for* TVS algorithm – (Solution-1b)

Predictive accuracy (in %age)

%age of Features Retained	Cylinder-Bands					Diabetes					Letter					Sonar					Waveform					Vehide					Glass					Arrhythmia					Average Predictive Accuracy
	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	Naive Bayes	J48	kNN	JRip	SVM	
100	72.22	57.78	74.44	65.19	81.67	76.3	73.83	70.18	76.04	77.34	97.3	99.49	99.88	99.3	97.17	67.79	71.15	86.54	73.08	75.96	80	75.08	73.62	79.2	86.68	44.8	72.46	69.86	68.56	74.35	48.6	66.82	70.56	68.69	56.07	62.39	64.38	52.88	70.8	70.13	**73.71**
95	72.41	57.78	74.81	67.41	82.04	76.56	73.96	65.76	73.57	77.47	96.99	99.35	99.83	99.23	97.08	68.27	70.19	85.1	73.56	78.37	80.04	75.28	73.4	79.88	86.58	44.68	73.17	69.27	64.66	72.34	50.47	67.29	77.1	66.36	51.87	63.05	65.27	52.65	69.69	70.35	**73.58**
90	72.41	57.78	75	66.85	82.04	76.56	73.96	65.76	73.57	77.47	96.78	99.06	99.64	99.01	96.93	68.75	70.67	85.1	75	77.88	79.98	75.5	74.08	79.54	86.78	44.33	73.17	69.39	67.26	71.28	50.47	67.29	77.1	66.36	51.87	61.95	63.5	51.77	68.58	69.91	**73.51**
85	72.41	57.78	75.93	66.3	82.59	76.17	73.57	65.76	73.96	76.69	96.62	99.06	99.55	99.03	96.93	68.27	74.04	86.06	74.04	77.88	80	75.86	74.64	79.7	86.76	45.27	73.17	70.57	65.84	71.51	47.66	70.09	77.1	62.15	51.87	60.84	61.95	51.33	70.13	70.35	**73.49**
80	72.59	57.78	76.11	66.3	82.96	76.17	73.57	65.76	73.96	76.69	96.61	98.91	99.44	98.89	96.95	71.15	76.44	85.58	72.12	79.81	79.98	76.16	74.72	80.38	86.76	44.44	71.75	72.46	69.15	71.75	47.66	70.09	77.1	62.15	51.87	60.4	64.38	51.77	69.91	71.02	**73.79**
75	71.67	57.78	76.48	66.85	82.22	76.17	73.57	65.76	73.96	76.69	96.61	98.91	99.44	98.89	96.95	71.63	76.44	84.62	73.56	79.33	79.96	76.22	75.32	79.7	86.7	43.85	71.63	73.29	67.73	71.28	46.26	72.9	73.36	60.28	51.87	59.51	64.82	51.11	68.81	70.8	**73.57**
70	71.3	57.78	76.11	68.15	80.37	74.87	72.4	67.45	71.88	74.48	96.89	98.64	99.04	98.45	96.94	71.15	74.04	83.65	71.15	75	79.96	75.98	75.22	79.1	86.74	45.04	71.28	72.34	68.68	70.57	46.26	72.9	73.36	60.28	51.87	61.28	63.27	50.22	69.47	72.12	**73.14**
65	71.85	56.67	77.04	67.78	79.81	74.87	72.4	67.45	71.88	74.48	96.36	98.3	98.7	98	95.94	71.15	74.04	82.69	74.04	77.4	80	76.02	76.28	79.26	86.92	44.56	69.86	71.63	66.9	70.21	47.66	71.5	72.9	62.62	51.4	61.95	61.95	49.34	68.81	71.46	**73.05**
60	72.04	56.67	77.04	70.19	80	74.87	72.53	66.93	72.4	74.48	96.38	97.88	97.99	97.89	95.94	68.75	71.15	82.69	77.88	75.48	80.08	76.36	77.38	79.48	86.9	44.8	70.21	72.81	67.02	69.5	47.66	71.5	72.9	62.62	51.4	59.96	61.95	50.22	67.26	70.13	**72.98**
55	69.81	56.67	77.04	64.26	80.19	74.87	72.53	66.93	72.4	74.48	94.75	97.59	97.16	97.37	95.94	65.38	72.12	79.81	76.44	73.08	80.1	76.3	77.5	79.62	86.8	46.45	70.69	71.75	65.13	68.32	50.93	74.3	74.77	64.49	51.4	59.73	63.27	50.22	70.58	68.14	**72.73**
50	70	56.67	76.3	66.85	80.74	74.87	72.53	66.93	72.4	74.48	94.75	97.59	97.16	97.37	95.94	65.38	71.63	84.13	74.52	74.04	80.06	76.36	78.08	80.02	86.86	46.45	70.69	71.75	65.13	68.32	50.93	74.3	74.77	64.49	51.4	59.73	63.27	49.56	65.49	69.47	**72.79**
45	70	56.67	77.41	65.19	79.81	75.13	72.53	67.84	72.79	75.39	95.94	96.89	96.1	96.68	95.94	67.31	72.12	81.25	75	73.56	80.36	76.96	78.7	80.06	86.8	48.23	71.99	71.04	67.73	67.73	50.93	74.3	74.77	64.49	51.4	60.62	63.72	49.78	69.47	68.58	**73.03**
40	70.19	56.67	78.89	65.93	80	75.13	72.53	67.84	72.79	75.39	95.94	95.93	94.96	96	95.94	67.79	75.96	79.33	72.6	72.6	80.2	77.06	77.82	79.16	86	48.58	71.75	70.57	67.85	66.67	46.73	66.36	72.9	67.76	46.73	61.5	62.61	48.23	68.36	69.25	**72.46**
35	69.44	56.67	81.48	61.85	76.48	74.61	72.53	67.84	72.4	75.26	95.94	95.94	95.87	95.95	95.94	64.9	76.92	78.37	71.63	75	80.16	74.78	75.56	78	84.12	50.24	70.21	67.85	67.38	54.96	46.73	66.36	72.9	67.76	46.73	62.17	64.38	47.79	68.14	68.36	**71.74**
30	69.63	56.67	80.93	56.3	76.48	74.61	72.53	67.84	72.4	75.2																					3.46	63.55	57.01	60.28	35.51	59.07	61.5	45.35	65.93	63.94	**69.27**
25	70.19	56.67	80	57.41	78.7	74.61	72.53	67.84	72.4	75.2																					3.46	63.55	57.01	60.28	35.51	59.29	61.95	44.03	65.93	63.27	**68.37**
20	70.19	56.67	80	61.11	78.7	67.19	67.84	67.32	67.19	65.																					5.98	54.67	47.2	52.8	35.51	61.5	61.95	46.24	66.15	63.27	**65.46**
15	70	56.67	80.56	60	77.96	67.19	67.84	67.32	67.19	65.																					5.98	54.67	47.2	52.8	35.51	63.05	61.5	52.65	65.04	61.73	**63.27**
10	74.63	57.78	74.26	60.37	77.96	65.1	65.1	65.1	65.1	65.																					5.51	35.51	35.51	35.51	35.51	63.05	54.2	52.21	65.04	61.5	**58.72**
5	61.48	57.78	54.81	57.78	76.85	65.1	65.1	65.1	65.1	65.																					5.51	35.51	35.51	35.51	35.51	60.18	49.34	47.12	61.5	61.5	**53.91**



- Total **800** experiments performed

Datasets & Experimental Setup – (Solution-1a & Solution-1b)

Selected Textual datasets characteristics

Textual Dataset	No. of Documents	No. of Features	No. of Distinct Classes	Description
MiniNewsGroups	800	27419	4	• Is a 10% subset of 20NewsGroups dataset, • Consider four equal sized categories - computer, politics, society and sport
Course-Cotrain	1051	13919	2	• Is a subset of 4Universities dataset and consists of web pages, • Consider two categories of pages - course and non-course
Trec05p-1	62499	12578	2	• Consists of e-mail documents, • Consider two categories of emails - spam and ham
SpamAssassin	3000	9351	2	• Consists of e-mail documents, • Consider two categories of emails - spam and ham

Selected Non-Textual dataset characteristics

Non-Textual Dataset	No. of Instances	No. of Features	No. of Distinct Classes	Description
Cylinder-bands	540	40	2	• Contains the process delay information of engraving printing for decision tree induction
Diabetes	768	9	2	• Consists of diagnostic measurements of patients • Consider two categories - has diabetes (YES) and not diabetes (NO)
Letter	20000	17	2	• Consists of black-and-white character image features • Identify English capital alphabet letter (from A to Z)
Sonar	208	61	2	• Contains signals information • Consider two bounced off categories of signals, namely "bounced off a metal cylinder" and "bounced off a roughly cylindrical rock"
Waveform	5000	41	3	• Contains 3 waves classes, which are produced by integrating 2 of 3 base waves
Vehicle	846	19	4	• Consists of silhouette features & consider four categories of vehicle
Glass	214	10	6	• Consists of oxide content & consider six categories of glass
Arrhythmia	452	280	13	• Consists of ECG records & consider thirteen categories of group • Consider two prediction categories of cardiac arrhythmia - presence of cardiac arrhythmia (YES) and absence of cardiac arrhythmia (NO)

Steps performed to preprocess the textual documents for applying the state-of-the-art and proposed algorithms:

- **Step-1:** Remove the structural content of the documents such as HTML or XML tags, sender and receiver fields in an e-mail document, links and etc.
- **Step-2:** Eliminate the pictures and e-mail attachments from the documents.
- **Step-3:** Tokenize the documents.
- **Step-4:** Remove the non-informative terms like stop-words from the contents.
- **Step-5:** Perform the terms stemming task.
- **Step-7:** Eliminate the low length terms whose length are less than or equal to 2.
- **Step-8:** Finally, generate the feature vectors representing document instances by computing the *term frequency–inverse document frequency* (tf-idf) weights.

Classifier	Function	Kernel Type	Epsilon	Tolerance	Exponent	Random Seed
SVM	SMO	Polynomial	1.0E-12	0.001	1	1

Why SVM classifier for evaluation process?

- The performance of SVM classifier is **better** as compared to other state-of-the-art classifiers such as KNN and Naïve Bayes [13].

Evaluation metric:

- **Predictive performance:** Precision, Recall, F-Measure (Uneven class distribution), and Accuracy (Symmetric dataset, where FP and FN are equal) [13].
- **Processing speed:** s (second)
- **Validation:** 10-fold cross-validation technique [28, 29]

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive (TP)	False Negative (FN)
	Class = No	False Positive (FP)	True Negative (TN)

- Precision = $\frac{TP}{TP + FP}$
- Recall = $\frac{TP}{TP + FN}$
- F-measure = $\frac{2 * (Recall * Precision)}{(Recall + Precision)}$
- Accuracy = $\frac{TP + TN}{TP + FP + FN + TN}$

Results & Discussion – (Solution-1a & 1b)

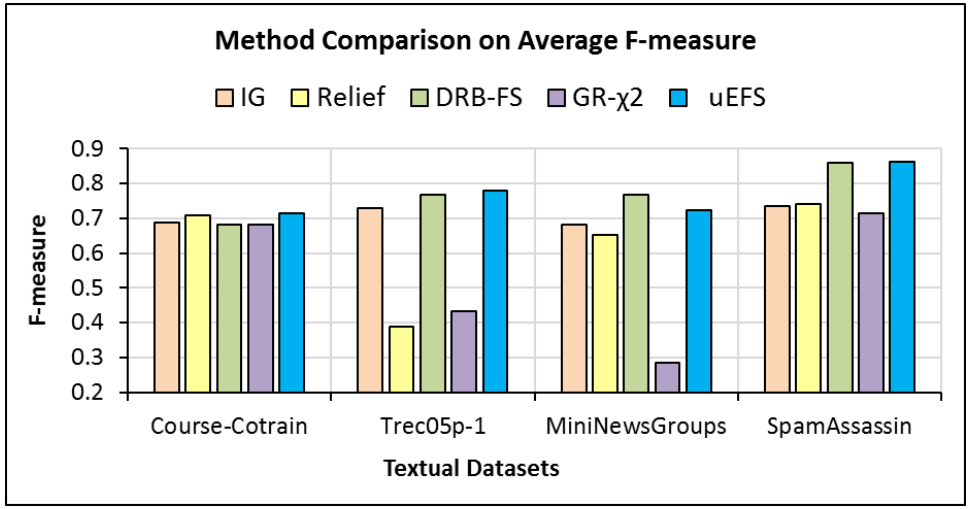


Figure: Comparisons of average F-measure of the uEFS with other state-of-the-art methods [13, 39, 40, 41]

Findings:

- Achieved on average **~7% increase** in **F-measure** as compared to baseline approach

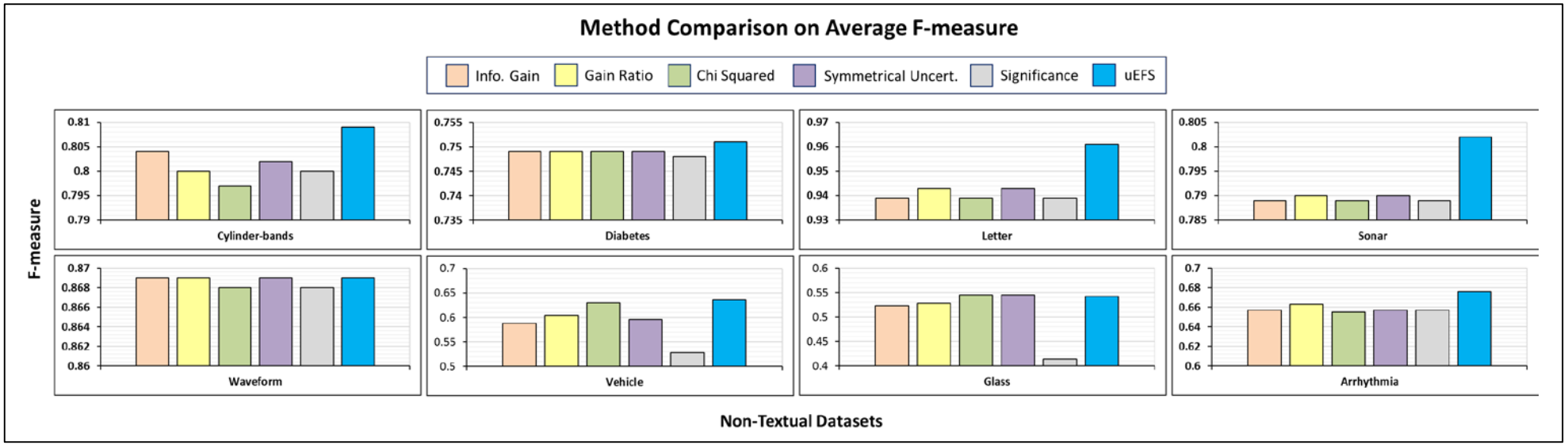
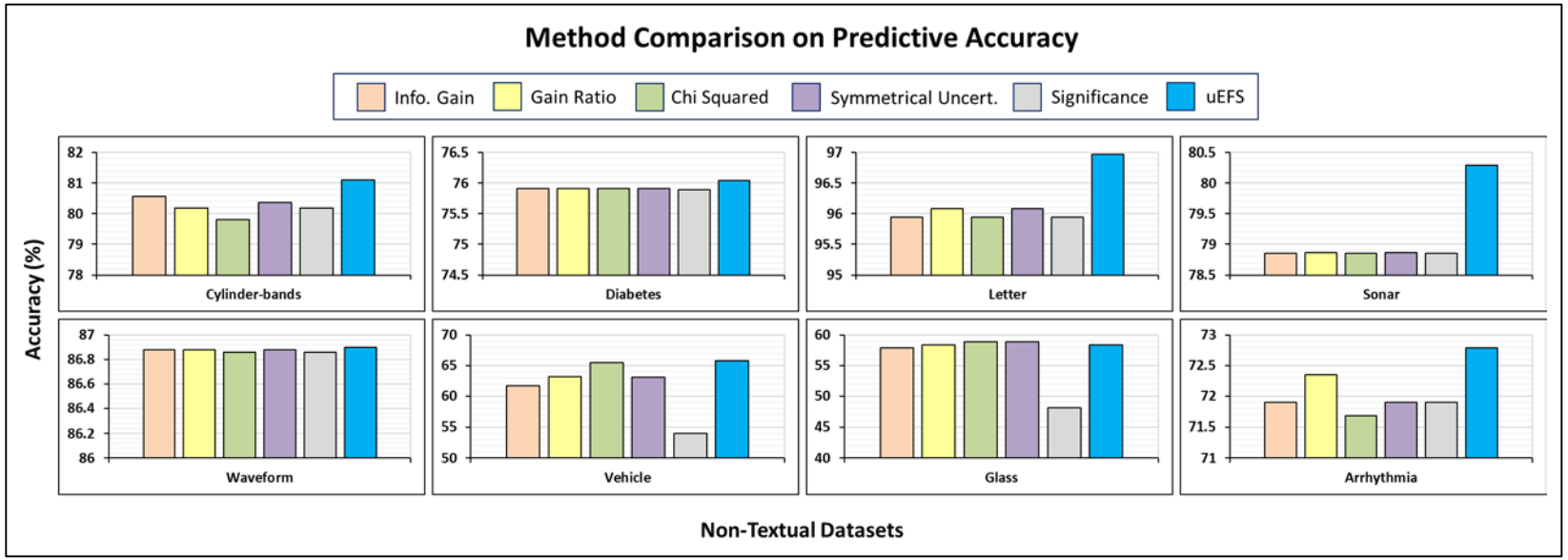


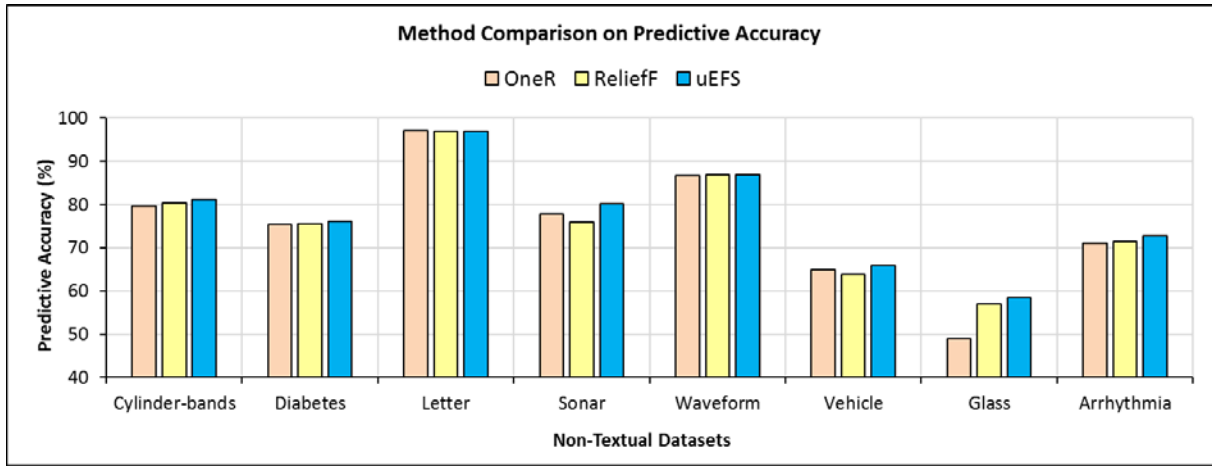
Figure: Comparisons of average F-measure of the uEFS with other state-of-the-art filter measures

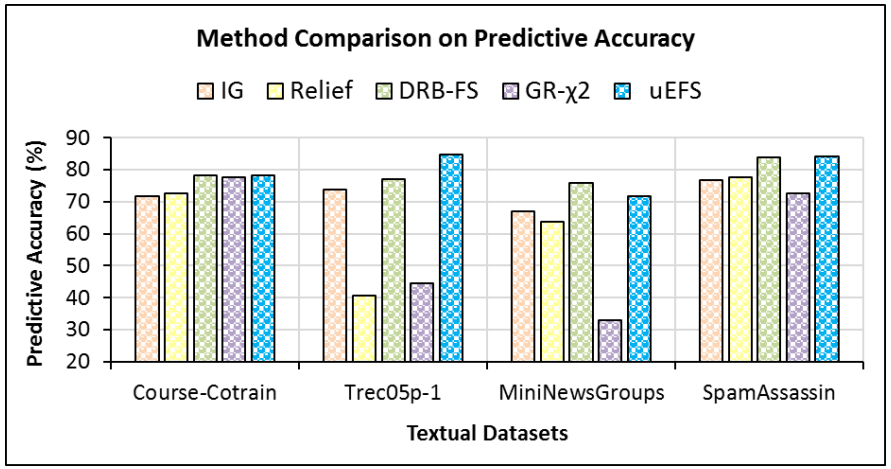
Results & Discussion – (Solution-1a & 1b)



Findings:

- Achieved on average **~5% increase** in **predictive accuracy** as compared to state- of-the-art filter methods





Figures: Comparisons of predictive accuracy (in %age) of the uEFS with other state-of-the-art filter methods

Results & Discussion – (Solutions-1a & 1b)

Table: Comparisons of predictive accuracy (%) with state-of-the-art filter measures

Non-Textual Dataset	Feature Selection Measures					Proposed Methodology	One-Sample T-Test	Paired-Samples T-Test
	Info. Gain	Gain Ratio	Chi Squared	Symmetrical Uncert.	Significance	uEFS	p {Sig.(2-tailed)}	p {Sig.(1-tailed)}
Cylinder-bands	80.56	80.19	79.81	80.37	80.19	81.11	0.002	0.010
Diabetes	75.91	75.91	75.91	75.91	75.89	76.04	0.000*	
Letter	95.94	96.08	95.94	96.08	95.94	96.97	0.000*	
Sonar	78.85	78.86	78.85	78.86	78.85	80.29	0.000*	
Waveform	86.88	86.88	86.86	86.88	86.86	86.9	0.005	
Vehicle	61.7	63.24	65.48	63.12	54.02	65.84	0.093	
Glass	57.94	58.41	58.88	58.88	48.13	58.41	0.400	
Arrhythmia	71.9	72.35	71.68	71.9	71.9	72.79	0.002	

One-Sample T-Test:

- Performed against each dataset
- Considered the uEFS value as a test value and feature selection measures' values as sample data.
  - For example, in case of **Cylinder-bands** dataset, 81.11 (value generated by the uEFS) is considered a test value, while 80.56, 80.19, 79.81, 80.37, 80.19 (values generated by *Info. Gain*, *Gain Ratio*, *Chi Squared*, *Symmetrical Uncert.*, *Significance*) are used as sample data.
- The mean feature selection measures score for Cylinder-band dataset **(M = 80.22, SD = 0.28)** was lower than the normal uEFS score of 81.11, a statistically significant mean difference of **0.89, 95% CI [0.54 to 1.23], t(4) = -7.141, p = .002**.

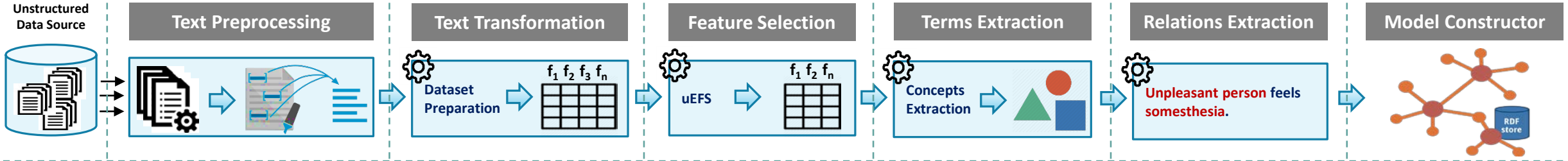
Paired-Samples T-Test

	State-of-the-art Filter-based Measures' Mean	Proposed Methodologyu EFS
Mean	75.970	77.294
Variance	164.664	144.659
Observations	8	8
Pearson Correlation	0.996	
Hypothesized Mean Difference	0	
df	7	
t Stat	-2.739	
P(T≤t) one-tail	0.014	
P(T≤t) two-tail	0.029	

Findings:

- It can be observed from the results of One-Sample T-test and Paired-Samples T-test that most of the significance (i.e. p) values are less than 0.05 (i.e. p < .05), which indicates that our proposed uEFS methodology results are statistically significantly different from state-of-the-art methods results.
- Variance value of the proposed methodology is decreased (indicates ➔ data points tend to be very close to the mean and more homogeneous).
- @Note: * This actually means that $p < 0.0005$. It does not mean that the significance level is actually zero.

Realization of Domain Knowledge Construction



Domains:
- **Diabetes**
- **Lung Cancer**

No. of documents:
- **19**

Tokenization: Chop the given text into pieces, called tokens.
Fiteration: Remove the non-informative terms (such as the, in, a, an, with, etc.).
Tagging: Assign each token with a parts-of-speech tag, such as noun, verb, etc.
Normalization: Identify the root/stem of a word. i.e. the words connected, connecting is stemmed to "connect".

Tool used: Rapid Miner Studio
Tokenization: English tokenizer
Fiteration: Stopword Removal
Tagging: Part-Of-Speech (POS) Tagger
Normalization: Porters Stemmer

blood	symptom	fertility	.....	specimen
0.009	0.002	0.002	.....	0.013
0.0	0.009	0.0	.....	0.0
0.0	0.007	0.0	.....	0.0
0.0	0.024	0.0	.....	0.0
0.024	0.007	0.006	.....	0.0

- Generate the feature vectors representing document instances

Technique used: Term Frequency – Inverse Document Frequency (TF-IDF)

Why TF-IDF?
- TF-IDF provides a good heuristic for determining likely candidate keywords [34].
- It is one of the best-known and most commonly used keyword extraction algorithms currently in use [35] when a document corpus is available.

symptom	feeling	blood	.....	disease
0.002	0.000	0.009	.....	0.004
0.009	0.001	0.0	.....	0.0
0.007	0.003	0.0	.....	0.0
0.024	0.001	0.0	.....	0.004
0.007	0.001	0.024	.....	0.0

Technique used: Proposed Univariate Ensemble Feature Selection (uEFS)
Step-1: Unified Features Scoring (UFS)
Step-2: Filtering features using Threshold Value Selection (TVS)

- Symptom
- Feeling
- Blood
- Unpleasant
- Person
- Negative
- Hurt
- disease

Process: Nouns, Verbs, Adjectives, and Adverbs Identification
Thesaurus used: Penn Treebank

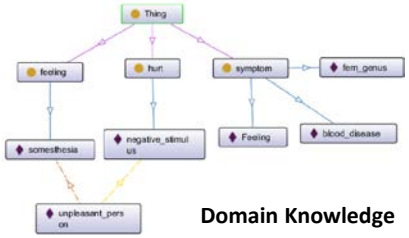
Why Penn Treebank?
- Treebank provides distinct coding for all classes of words having distinct grammatical behavior [36].

- Unpleasant_person feels somesthesia.
- Unpleasant_person has negative_stimulus.
- Blood_disease is a symptom.

No of Relations = **1550**

Technique used: Lexical Chaining
Thesaurus used: Princeton's WordNet
Process: Hypernyms Identification
Keep original tokens: True
Multiple meanings per word policy: Take all meanings per token
Multiple synset words: Take only first synset word
Validation: Domain Expert

Why Lexical Chains?
- A well known technique for text connectivity [37] that locate terms and their sequence in accurate manner [34].



Domain Knowledge

Language used: Attempto Controlled English (ACE)
Editor used: ACE View

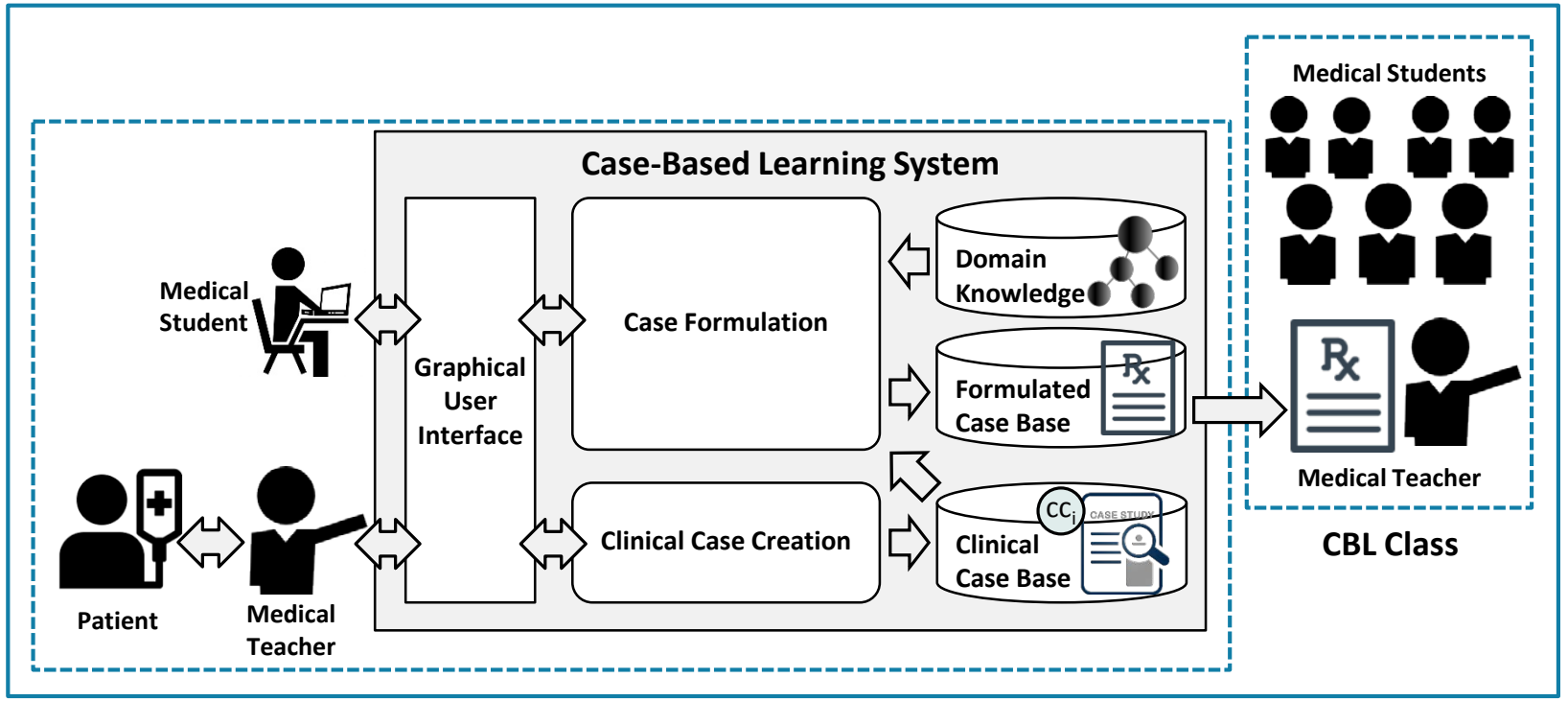
Why ACE? [38]
- A logic-based knowledge representation language
- Uses the syntax of a subset of English
- Provides automatic and unambiguous translation of text into first-order logic

Proposed *Case-Based Learning* (CBL) Approach – (Solution-2)

Traditional CBL Approach

CBL Class

Proposed CBL Approach



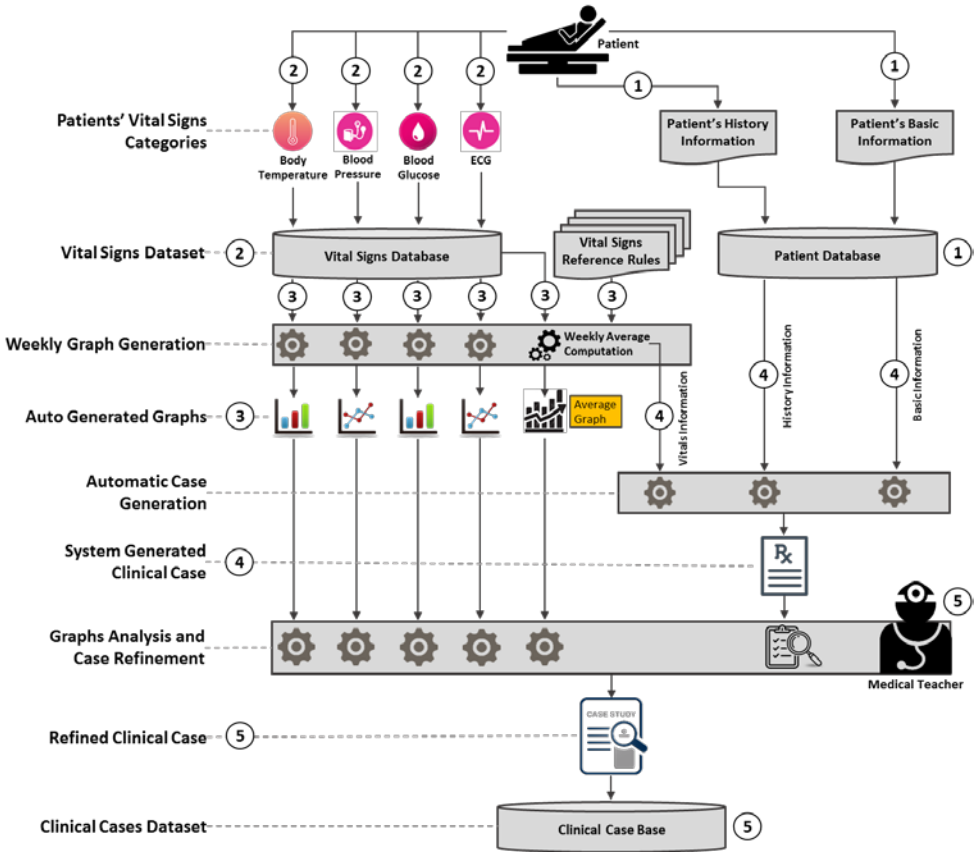
Highlight of the proposed idea

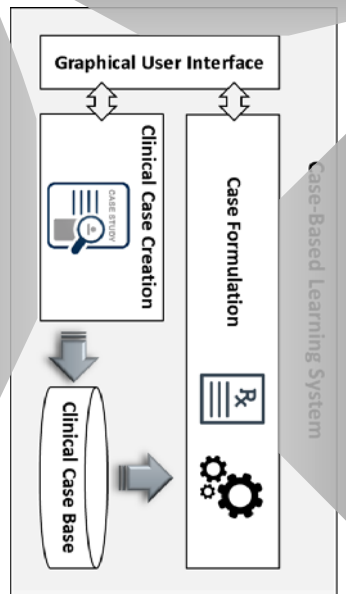
- Enables the medical teacher to create real-world CBL cases for their students, review the students' solutions, and to give feedback and opinions to their students.
- Facilitates the medical students to do the CBL rehearsal before attending actual CBL class.

Maqbool Ali et al., IoTFLiP: IoT-based Flip Learning Platform for Medical Education, *Digital Communications and Networks*, vol. 3, pp.188–194, 2017.
Maqbool Ali et al., iCBLS: An interactive case-based learning system for medical education, *International journal of medical informatics*, vol. 109, pp. 55-69, 2018.

Proposed *Clinical Case Creation* and *Formulation* Techniques – (Solution-2)

1. Maqbool Ali et al., An IoT-based learning methodology for medical students' education, *Korean Intellectual Property Office*, Registration No.(Date) 1018088360000 (2017.12.07).
2. Maqbool Ali et al., An IoT-based CBL Methodology to Create Real-world Clinical Cases for Medical Education, In *ICTC 2017*, pp.1037-1040, IEEE, 2017.
3. Maqbool Ali et al., An Interactive Case-Based Flip Learning Tool for Medical Education, In *ICOST 2015*, pp.355-360, 2015.
4. Maqbool Ali et al., iCBLS: An interactive case-based learning system for medical education. *International journal of medical informatics*, vol. 109, pp. 55-69, 2018.





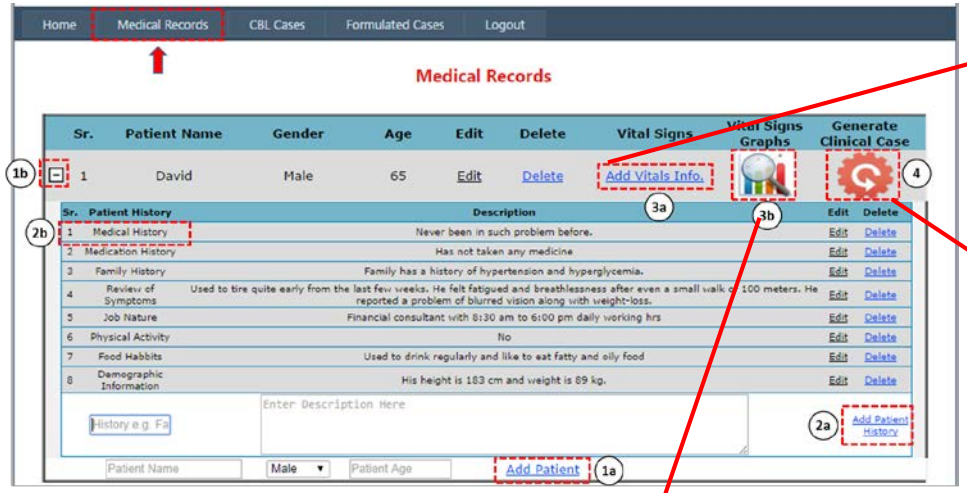
Case Formulation:

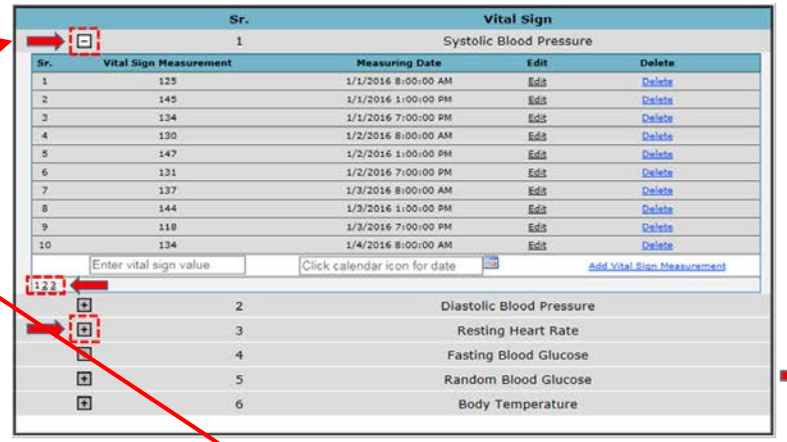
- Formulating a clinical case involves constructing appropriate interpretations about a patient's problem to *create* a significant medical story within the context of his or her life [30].

Algorithm 2: Case Formulation($D = CC[Ref.$**Algorithm 1**$]$)

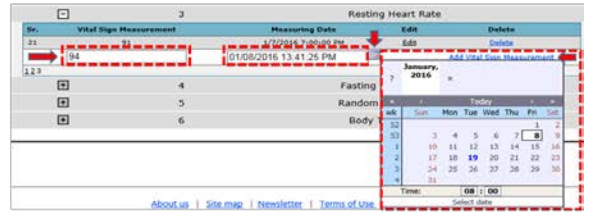
Data: $D = CC$: Input dataset (clinical case)
Result: CF – Case Formulation
if $Verify(D)$ **then**
  /* For creating the medical charts, add the components of medical charts e.g. presenting complaints, previous medications for D cases */;
  $MCC \leftarrow addMedicalChartComponent(D)$: $D = mcc_1, mcc_2, mcc_3, ..., mcc_n$;
  for $\forall\, mcc_m \in D$ **do**
    /* Add observations e.g. felt fatigue, breathlessness of each chart component mcc_m */;
    $Obs \leftarrow addObservations(mcc_m)$;
  end
  /* Case formulation includes information of medical charts component and observations */;
  $CF \leftarrow caseFormulation(MCC, Obs)$;
  return CF : *case formulation*
else
  $Error(message)$;
end

Realization of the *Clinical Case Creation approach* – (Solution-2)

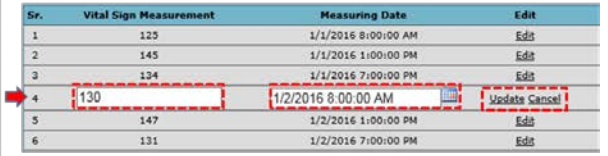




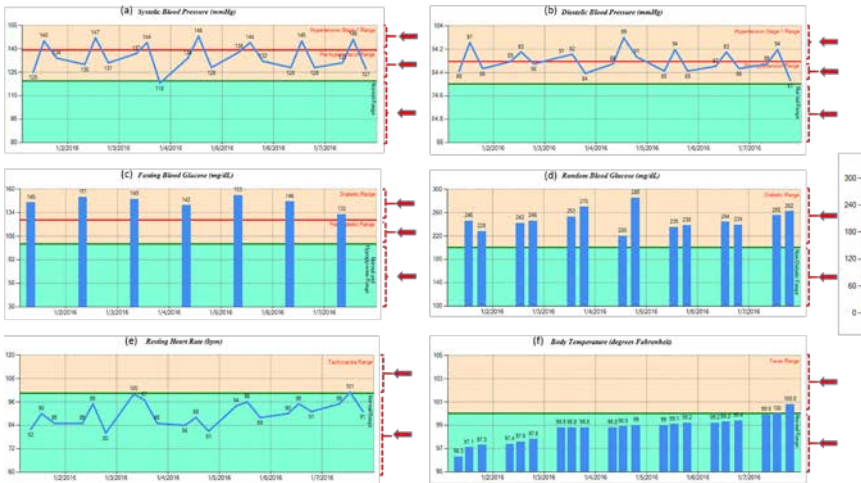
(a) Patient's vital signs measurement information

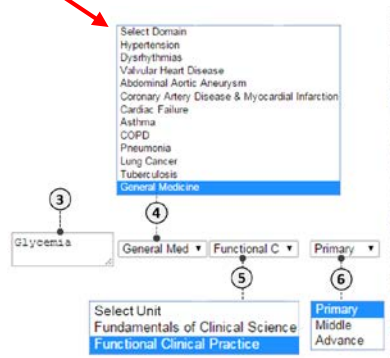


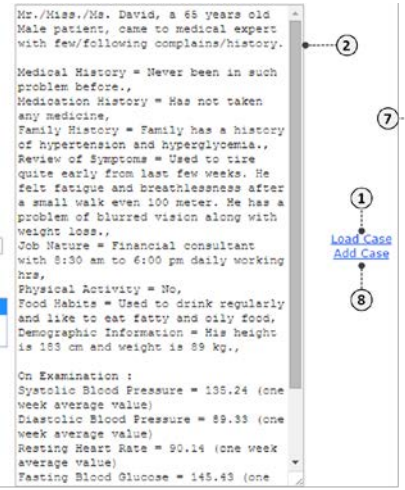
(b) Adding vital sign measurement value and date

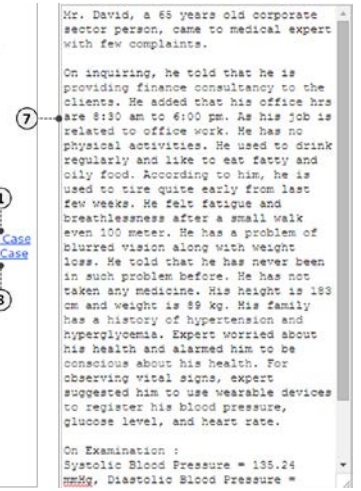


(c) Modifying vital sign measurement value and date









Maqbool Ali et al., An IoT-based CBL Methodology to Create Real-world Clinical Cases for Medical Education, In *ICTC 2017*, pp.1037-1040, IEEE, 2017.
Maqbool Ali et al., iCBLS: An interactive case-based learning system for medical education. *International journal of medical informatics*, vol. 109, pp. 55-69, 2018.

Evaluation Setup – (Solution-2)

Evaluation

- For holistic understanding, the proposed system is evaluated in *heterogeneous environments* by involving multiple stakeholders and using multiple methods such as (1) *quantitative methods* (e.g. surveys) and (2) *qualitative methods* (e.g. interviews and focus groups) under the umbrella of the CIPP (context/input/process/product) model.

Reason for choosing CIPP model:

- Discussion-based learning in a small-group, like CBL, is considered to be a complex system [31] due to having multiple interaction of students and exchanging information with each other [32, 33].
- For evaluation of complex systems, the CIPP model is most widely used and is considered as a powerful approach [32].

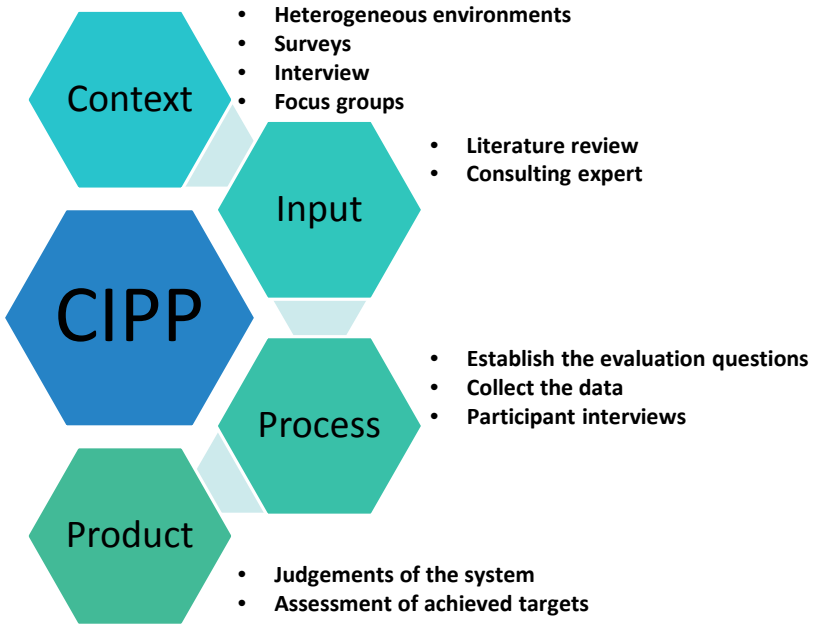


Figure: CIPP elements and tasks performed [32].

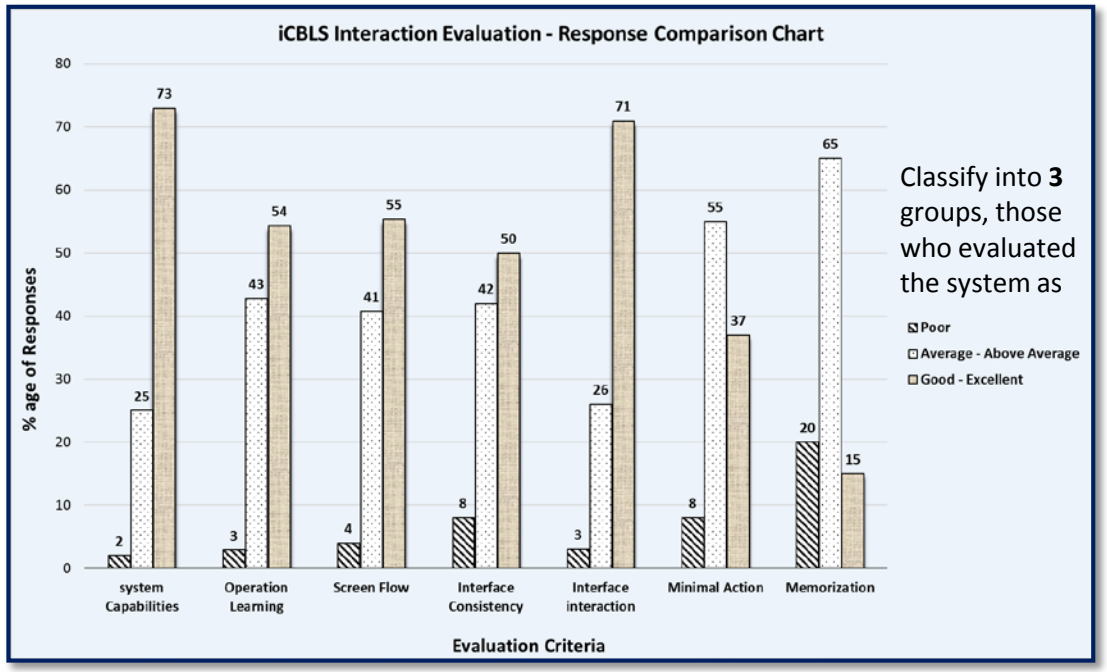
Evaluation Setup

Evaluation Criteria	Environment-I (Users Interaction Evaluation)	Environment-II (Learning Effectiveness Evaluation)
Primary hypothesis	Flexible and easy to learn	System appropriateness with respect to students' learning
Secondary hypothesis	Minimum memory load and efficiency (minimum actions required)	System suitability with respect to students' level and user friendly system
Variables	System capability, Operation learning, Screen flow, Interface consistency, Interface interaction, Minimal action, Memorization	Appropriate for group learning, Appropriate for solo learning, Useful for improving clinical skills, Performing tasks straightforward
Options and weightages set for each question	Excellent (10), Good (8), Above Average (6), Aver- age (4), Poor (2)	Five options from 1 to 5 representing poor to excellent and quantified in multiple of 20
Survey method	Google docs (Online), 1-on-1	Google docs (Online), 1-on-1, small groups at the hospital
Number of users	209 (different years students and professionals)	

Results & Discussion – (Users Interaction Evaluation) – (Solution-2)

Summarized response with respect to categories results

Evaluation Criteria		Sub-categories Response	Categories Response	
Categories	**Sub-categories**	**(out of 10)**	**(Average)**	**(%)**
System Capability	System reliability	7.5555	7.8148	78.15
	Designed for all levels of users	8.0740		
Operation Learning	Learning to operate the system	7.2963	7.2037	72.04
	Reasonable Data grouping for easy learning	7.1111		
Screen Flow	Reading characters on the screen	6.9629	7.0555	70.56
	Organization of information	7.1481		
Interface Consistency	Consistency across the label format and location	7.1111	6.6851	66.85
	Consistent symbols for graphic data standard	6.2592		
Interface Interaction	Flexible data entry design	8.0000	8.1481	81.48
	Zooming for display expansion	8.2962		
Minimal Action	Wizard-based information management	6.7407	6.0185	60.19
	Provision of default values	5.2962		
Memorization	Highlighted selected information	4.8148	4.8148	48.15



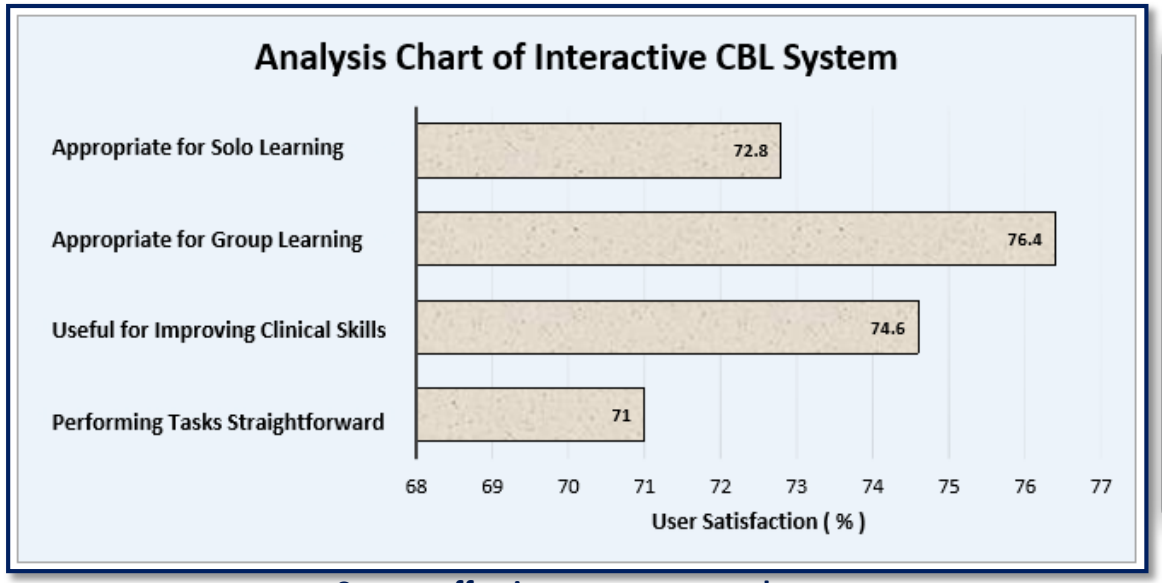
Findings:

- Interaction of the system through the interface was generally valued by the users
- Users were *quite satisfied* with the *system capabilities*, *operating learning*, *screen flow*, and *interface interaction*, which were greater than **70%**.

Findings:

- The confidence on the *system capabilities* and the *interface interaction* was measured as about **70%** from all users.
- Approximately **50%** of users considered the *interface consistency*, *screen flow* and *operation learning* aspect as an appealing factor.

Results & Discussion – (Learning Effectiveness Evaluation) – (Solution-2)



System effectiveness summary chart

Findings:

- Users were quite satisfied with the system appropriateness for group as well as solo learning, system usefulness with respect to enhancing clinical skills, and user friendliness of the system, which were **greater than 70%**.
- The system was also evaluated to check *suitability* and *appropriateness* for different course-year levels of medical students. The system achieved **votes for year-levels 2 or 3** that showed confidence on system suitability for these students, which is the stage where students begin to do placements at hospitals.

Open-ended Survey Question for Learning Effectiveness Evaluation

#	Open-ended Survey Questions
1	What did you like most about the computer-based tutorial preparation module?
2	What did you like least about the computer-based tutorial preparation module?
3	Are there any areas where you think the Case-Based Learning tutorial program can improve?

Findings:

- System encouraged the students to be active learners, and to use logic to think and learn with real-world cases
- **Key phrases** from answers were '*self-learning*', '*independent thinking*', '*gaining more professional knowledge*', '*distance learning*', '*senior level education*', '*tutor engagement*', and '*improvement of feedback interface*'.

Conclusion

This thesis contributes to

1. An *efficient* and *comprehensive* ensemble-based feature selection methodology
   - Proposed a flexible approach (UFS) for incorporating state-of-the-art univariate filter measures for feature ranking
   - Proposed an efficient approach (TVS) for selecting a cut-off value for the threshold in order to select a subset of features
   - Performed extensive experimentation for the proof-of-concept for the aforementioned techniques.
     - Achieved on average **~7% increase** in *F-measure* as compared to baseline approach
     - Achieved on average **~5% increase** in *Predictive Accuracy* as compared to state-of-the-art methods.
2. An *interactive* and *effective* Case-Based Learning (CBL) approach for medical education
   - Introduce a real-world clinical case creation and case formulation techniques
   - The proposed CBL approach achieves a success rate of more than **70%** for *students' interaction*, *group learning*, *solo learning*, and *improving clinical skills*.

Uniqueness

- A *comprehensive* and *flexible* feature selection methodology based on an ensemble of univariate filter measures.
- An *effective* CBL approach using real-world clinical case creation and case formulation support.

Future Work

Applications

uEFS methodology contributes in feature selection, which is the key step in most of decision support system

- Data-driven knowledge acquisition system[1]
- Case-based learning system[2]
- Clinical decision support system

Limitation

- Only univariate filter measures are considered in the proposed methodology
- This methodology does not evaluate the suitability of a measure, a precision
- On average, the proposed methodology takes 0.37 sec more time than state-of-the-art filter measures

Future work

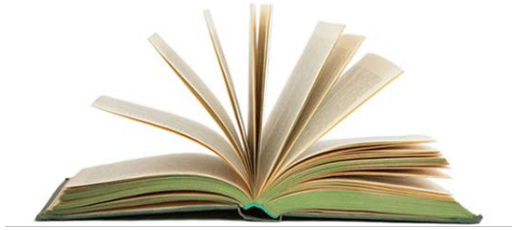
- Extend the methodology for incorporating multi-variate measures
- Investigate the application of fuzzy-logic for determining the cut-off threshold value
- Extend the CBL towards QA-based learning environment

[1]Maqbool Ali et al., A data-driven knowledge acquisition system: An end-to-end knowledge engineering process for generating production rules, *IEEE Access*, vol. 6, pp. 15587-15607, 2018.

[2]Maqbool Ali et al., iCBLS: An interactive case-based learning system for medical education. *International journal of medical informatics*, vol. 109, pp. 55-69, 2018.

Publications

- **Published papers**
  - **Patents (03)**
    - Three Korean
  - **SCI / SCIE Journals (06)**
    - **SCI (01)**
    - **SCIE (01)**
    - **Co-author (04)**
  - **Non-SCI Journal (02)**
    - ESCI (01)
    - Co-author (01)
  - **Conferences (14)**
    - International (06)
    - Domestic (03)
    - Co-author (05)



Publication

Total Publications (25)

First Author Publications (15)

❖ **Paper in progress**

- SCIE Journal (01)
  - Maqbool Ali et. al.. "An efficient and comprehensive ensemble-based feature selection methodology to select informative features from an input dataset". PLOS ONE. Under review, 2018.

References

[1] Jill Elizabeth Thistlethwaite, David Davies, Samilia Ekeocha, Jane M Kidd, Colin MacDougall, Paul Matthews, Judith Purkis, and Diane Clay. The effectiveness of case-based learning in health professional education. a beme systematic review: Beme guide no. 23. Medical teacher, 34(6):e421–e444, 2012.

[2] Kaitlyn Brown, Mary Commandant, Adi Kartolo, C Rowed, A Stanek, H Sultan, K Tool, and V Wininger. Case based learning teaching methodology in undergraduate health sciences. Inter disciplin. J. Health Sci, 2(2):47–65, 2011.

[3] Sharon R Stewart and Lori S Gonzalez. Instruction in professional issues using a cooperative learning, case study approach. Communication Disorders Quarterly, 27(3):159–172, 2006.

[4] Baitule, P., & Chole, V. (2014). A review on improved text mining approach for conversion of unstructured to structured text'. International Journal of Computer Science and Mobile Computing, 3(12), 156-159.

[5] Joseph, S., Mugauri, C., & Sumathy, S. (2017, November). Sentiment analysis of feature ranking methods for classification accuracy. In IOP Conference Series: Materials Science and Engineering (Vol. 263, No. 4, p. 042011). IOP Publishing.

[6] Dhote, Y., Agrawal, S., & Deen, A. J. (2015, December). A survey on feature selection techniques for internet traffic classification. In Computational Intelligence and Communication Networks (CICN), 2015 International Conference on (pp. 1375-1380). IEEE.

[7] Rusu, O. et al. Converting unstructured and semi-structured data into knowledge. In Roedunet International Conference (RoEduNet), 2013 11th (pp. 1-4). IEEE.

[8] Deng, K. (1998). OMEGA: On-line memory-based general purpose system classifier (Doctoral dissertation, Carnegie Mellon University).

[9] Tuv, E., Borisov, A., & Torkkola, K. (2006, July). Feature selection using ensemble based ranking against artificial contrasts. In Neural Networks, 2006. IJCNN'06. International Joint Conference on (pp. 2181-2186). IEEE.

[10] Sarkar, C., Cooley, S., & Srivastava, J. (2014). Robust feature selection technique using rank aggregation. Applied Artificial Intelligence, 28(3), 243-257.

[11] Osanaiye, O., Cai, H., Choo, K. K. R., Dehghantanha, A., Xu, Z., & Dlodlo, M. (2016). Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. EURASIP Journal on Wireless Communications and Networking, 2016(1), 130.

[12] Ali, M., Han, S. C., Bilal, H. S. M., Lee, S., Kang, M. J. Y., Kang, B. H., ... & Amin, M. B. (2018). iCBLS: An interactive case-based learning system for medical education. International journal of medical informatics, 109, 55-69.

[13] Sadeghi, S., & Beigy, H. (2013). A new ensemble method for feature ranking in text mining. International Journal on Artificial Intelligence Tools, 22(03), 1350010.

[14] Altidor W. Stability analysis of feature selection approaches with low quality data. Florida Atlantic Uni.; 2011.

[15] Stoean R, Gorunescu F. A survey on feature ranking by means of evolutionary computation. Annals of the University of Craiova-Mathematics and Computer Science Series. 2013;40(1):100-105.

[16] Doraisamy S, Golzari S, Mohd N, Sulaiman MN, Udzir NI. A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music. In: ISMIR; 2008. p. 331-336.

[17] Liu, C., Wang, W., Zhao, Q., Shen, X., & Konan, M. (2017). A new feature selection method based on a validity index of feature subset. Pattern Recognition Letters, 92, 1-8.

[18] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-1182.

[19] Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1).

[20] Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. Journal of Information Science, 43(1), 25-38.

References

[21] UTMB. Design a case, university of texas medical branch - utmb. http://www.designacase.org/default. aspx, Accessed: 2017-01-10.

[22] UNM. Extension for community healthcare outcomes - echo, the university of new mexico. http: //echo.unm.edu/, Accessed: 2016-12-16.

[23] Lih-Shyang Chen, Yuh-Ming Cheng, Weng Sheng-Feng, Chen Yong-Guo, and Chyi-Her Lin. Applications of a time sequence mechanism in the simulation cases of a web-based medical problem-based learning system. Journal of Educational Technology & Society, 12(1):149, 2009.

[24] Sharma, A. and Dey, S. (2012). Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications, 3, pp.15-20.

[25] Ali, S.I. and Shahzad, W. (2012). A feature subset selection method based on symmetric uncertainty and ant colony optimization. In Emerging Technologies (ICET), 2012 International Conference on (pp. 1-6). IEEE.

[26] Ahmad, A. and Dey, L. (2005). A feature selection technique for classificatory analysis. Pattern Recognition Letters, 26(1), pp.43-56.

[27] Belanche, L. A., & González, F. F. (2011). Review and evaluation of feature selection algorithms in synthetic problems. arXiv preprint arXiv:1101.2320.

[28] Prati, R. C. (2012, June). Combining feature ranking algorithms through rank aggregation. In Neural Networks (IJCNN), The 2012 International Joint Conference on (pp. 1-8). IEEE.

[29] McLachlan, G., Do, K. A., & Ambroise, C. (2005). Analyzing microarray gene expression data (Vol. 422). John Wiley & Sons.

[30] M.D. Adam Blatner. The art of case formulation. http://www.blatner.com/adam/psyntbk/formulation. html, 2006. Accessed: 2016-12-18.

[31] S. Mennin, Small-group problem-based learning as a complex adaptive system, Teaching and Teacher Education 23 (3) (2007) 303–313.

[32] A. W. Frye, P. A. Hemmer, Program evaluation models and related theories: Amee guide no. 67, Medical teacher 34 (5) (2012) e288–e299.

[33] S. Mennin, Teaching, learning, complexity and health professions education, J Int Assoc Med Sci Educat 20 (2010) 162–165.

[34] Lott, B. (2012). Survey of keyword extraction techniques. *UNM Education*, *50*.

[35] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, *60*(5), 503-520.

[36] Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. In *Treebanks* (pp. 5-22). Springer, Dordrecht.

[37] Ghosh, Avishikta. "Bengali Text Summarization using Singular Value Decomposition." PhD diss., 2014.

[38] Kuhn, T. (2009). *Controlled English for knowledge representation* (Doctoral dissertation, Doctoral thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, Switzerland, to appear).

[39] Lutu, P. E., & Engelbrecht, A. P. (2010). A decision rule-based method for feature selection in predictive data mining. Expert Systems with Applications, 37(1), 602-609.

[40] Makrehchi, M., “Feature ranking for text classifiers”, Ph.D. Thesis, Department of Electrical and Computer Engineering, University of Waterloo, (2007).

[41] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine Learning Proceedings 1992 (pp. 249-256).

Thank you
for your attention

Q & A ?