Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbour classifier

Loganathan Meenachi ✉, *Srinivasan Ramakrishnan*

Department of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India
✉ *E-mail: lmeenachi@gmail.com*

Published in Healthcare Technology Letters; Received on 13th June 2018; Accepted on 29th June 2018

Cancer is one of the deadly diseases of human life. The patient may likely to survive if the disease is diagnosed in its early stages. In this Letter, the authors propose a genetic search fuzzy rough (GSFR) feature selection algorithm, which is hybridised using the evolutionary sequential genetic search technique and fuzzy rough set to select features. The genetic operator's selection, crossover and mutation are applied to generate the subset of features from dataset. The generated subset is subjected to the evaluation with the modified dependency function of the fuzzy rough set using positive and boundary regions, which act as a fitness function. The generation and evaluation of the subset of features continue until the best subset is arrived at to develop the classification model. Selected features are applied to the different classifiers, from the classifiers fuzzy-rough nearest neighbour (FRNN) classifier, which outperforms in terms of classification accuracy and computation time. Hence, the FRNN is applied for performance analysis of existing feature selection algorithms against the proposed GSFR feature selection algorithm. The result generated from the proposed GSFR feature selection algorithm proved to be precise when compared to other feature selection algorithms.

1. Introduction: The computerised classification of data for diagnosis of diseases plays a pivotal role in biomedical field. The cancer is one of the deadly diseases. For every disease or ailment for that matter, early the diagnosis, lesser the mortality rate. Various computational diagnostic techniques are available to classify the malady. The clinical dataset or the dataset made known from machine learning repository form the basis to predict the cancer data and the predicted data for early clinical decision with higher reliability. From the large dataset, the classification of cancer stage of a patient is diagnosed by taking into account the informative features [1].

The feature selection is the process where the irrelevant, redundant features are removed from the original features [2]. The feature selection methods are categorised as filter, wrapper and hybrid methods [3]. In filter method, the features are considered individually by ranking and selection of features, which is either by forward selection or by backward elimination. It never receives feedback from any of the learning algorithm. The next method is wrapper method wherein the subset of features are taken and evaluated through the induction of the learning algorithm. The last method is hybrid method, which is a combination of both.

The model generated from the classification produces better accuracy with lesser computation time through the informative features from the dataset [1–4]. There are numerous research done in cancer data classification. Onan [5] proposed an approach of fuzzy-rough nearest neighbour (FRNN) classifier which is combined with consistency based subset evaluation and fuzzy rough instance selection algorithms with a view to generating the model with minimum features and instances that provide accurate classification similar to that of the whole dataset.

Jain *et al.* [6] proposed a hybrid model for cancer classification by hybridising correlation-based feature selection (CFS) and improved-binary particle swarm optimisation. This hybrid model selects the features to classify the binary and multi-class cancers using Naive–Bayes classifier to gain better accuracy. Korfiatis *et al.* [7] proposed the diagnosis method based on new wrapper feature selection algorithmic scheme for the primary and secondary polycythaemia. The novel wrapper feature selection algorithm is used from the local maximisation which acts as an initial dataset. Further, it refines and produces the subset which is applied for classification to achieve better accuracy. Narendar Reddy and Ravi [8] proposed two classification techniques, one is based on kernel principal component analysis and another based on quartile regression for effective classification.

Chuang *et al.* [9] have proposed the K-nearest neighbour (K-NN) classification with the lowest error rate by applying the reduced features. This is carried out through the hybrid method of feature selection by involving correlation-based feature selection and the Taguchi chaotic binary particle swarm optimisation. Baitharua and Panib [10] proposed an approach for discovering the hidden patterns using the decision algorithm J48 to classify the diseases in a different scenario with better accuracy and lesser computation time in the early stage.

Kumar and Rath [11] proposed a method for classifying the microarray dataset by reducing the features using the t-test as a feature selection method. The reduced feature is classified with KNN and fuzzy KNN algorithms. Comparing the performance measures, the fuzzy K-NN model entailed a better model than the K-NN. Zhou and Dickerson [12] proposed a novel class-dependent feature selection technique to select the desirable features. Support vector machine (SVM) combined with fuzzy KNN helps to classify the cancer using class-dependent features. Qu *et al.* [13] presented the FRNN classifiers and its merits. This paper highlights the accuracy of the FRNN based on the similarity between the objects in the training and test dataset. Wei *et al.* [14] proposed a rough set theory based classification of different types of cancers by analysing implicit hypercuboids.

Maa and Xia [15] have proposed a method where the whole dataset is divided into multiple tribes and generated an optimal subset to produce the accurate pattern classification. Jensen and Cornelis [16] proposed the nearest neighbour (NN) algorithm using both lower and upper approximations from the fuzzy rough set to classify the uncertainty in both fuzzy and rough set. While comparing the other NN classifiers it outperforms the other classification in terms of results.

Based on the related works discussed, the feature selection plays a major role in the data classification. The existing techniques also perform feature selection, even then few limitations occur as follows: the classification cannot provide a significant computation time when a different variety of dataset is used, the highly correlated features which are selected for classification results fail to bring about the important feature for classification, as it produces

a higher error rate with a limited classification accuracy [5–8]. The above said limitations are addressed in this Letter by using metaheuristic search for probing different characteristic dataset. The function used for subset evaluation addresses inadequate features to reduce the error rate and to obtain the stopping condition for subset search. The contributions of this Letter are

- Evolutionary sequential genetic search technique and fuzzy rough set are hybridised in this Letter to devise the genetic search fuzzy rough (GSFR) feature selection algorithm.
- Considering both positive and boundary regions of the fuzzy rough set, modified dependency function is recommended.
- The devised modified dependency function is applied as a fitness function in the proposed GSFR feature selection algorithm.

The proposed Letter is structured as follows: Section 2 presents the synopsis of rough set theory, fuzzy rough set, and evolutionary sequential genetic algorithm, FRNN algorithm, dataset description and evaluation parameters. Section 3 talks about the proposed method. Section 4 interprets the result and discussions. Finally, the conclusion part of this proposed method is presented in Section 5.

2. Theoretical background

2.1. Rough set theory: The rough set theory is a tool used to determine the data dependencies and to reduce the dimensions of the dataset [14]. The dataset with discrete attribute values is used to find out the subset of attributes with minimum loss of information. Let I be the Information system it is represented as

$$I = (U, A) \tag{1}$$

where U is the non-empty set of universe and A is the non-empty finite set of attributes such that a: $U \rightarrow V_a$ for every $a \in A$. V_a set of values for the attribute a. The attribute set is separated into two non-empty disjoint subsets C and D, where C is the conditional set and D is the decision set. The indiscernibility relation IND (P) with $P \subseteq A$ [17], the equivalence relation is defined as follows:

$$\mathrm{IND}(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \tag{2}$$

where P is the subset of attribute set, x and y are objects. The set of equivalence classes is denoted by U/IND (P). The information in P is approximated by constructing lower and upper approximations of P [14]

$$\underline{P}X = \{x | [x]_P \subseteq X\} \tag{3}$$

$$\bar{P}X = \{x | [x]_P \cap X \neq \emptyset\} \tag{4}$$

The positive region can be defined as

$$\mathrm{POS}_P(Q) = U_{X \in U/Q} \underline{P}X \tag{5}$$

where P and Q are the equivalence relations on U. Once the positive region is derived, the degree of dependency can be obtained

$$k = \frac{|\mathrm{POS}_P(Q)|}{|U|} \tag{6}$$

The attributes are reduced by the equivalence relations generated by the set of attributes. The prediction of the reduced attribute remains the same.

2.2. Fuzzy rough set: It denotes the generalisation of rough set and which is the approximation of a fuzzy set over the crisp dataset [9]. Fuzzy rough set operates on the real valued attributes, whereas the rough set operates on discrete values only. The fuzziness is integrated into the rough set so that the membership value of fuzzy set removes the roughness of the boundary region, which ranges from 0 to 1.

The fuzzy rough set approach operates on the real valued attribute in order to exercise the fuzziness as vagueness exists in the real valued dataset. The fuzzy sets $F = \{F_1, F_2, \ldots, F_n\}$ which is expressed from an equivalence class. F_i is a fuzzy set where $i \in \{1, 2, \ldots, n\}$, $\mu_x(x)$ is the membership where x belongs to the fuzzy set X executed on the universe U.

The lower and upper approximations are given as

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\{1 - \mu_{Fi}(x), \mu_X(x)\} \forall i \tag{7}$$

$$\mu_{\bar{P}X}(F_i) = \sup_x \min\{\mu_{Fi}(x), \mu_X(x)\} \forall i \tag{8}$$

The domain F_i is defined as

$$\gamma_C(D) = \frac{\sum_{x \in U} \mu_{\mathrm{POS}_P(Q)}(x)}{|U|} \tag{9}$$

where $\mu_{\mathrm{POS}_C(D)}(x)$ is the positive region and U is a non-empty finite set of object.

The boundary region of F_i is defined as

$$\mu_{\mathrm{BND}_C(D)}(x) = \mu\bar{P}X(x) - \mu\underline{P}X(x) \tag{10}$$

where $\mu_{\mathrm{BND}_C(D)}(x)$ is the boundary region, $\mu\underline{P}X(x)$ is the lower approximation, $\mu\bar{P}X(x)$ is the upper approximation.

2.3. Evolutionary sequential genetic search algorithm: A genetic algorithm is a random evolutionary search technique used to find out vector $\boldsymbol{x}$ (string). The evolutionary sequential genetic algorithm generates the new population of string from the existing population. The generated string is an encoded version of a tentative solution. The evaluation function is used as the fitness measure in every string by indicating its suitability to the problem. In order to generate new stings, the genetic algorithm applies genetic operators such as selection, crossover and mutation on random population [15].

2.4. FRNN classifier: It captures the fuzzy and rough uncertainties in the dataset [16]. The uncertainty of fuzzy and rough set is due to intersecting classes and inadequate features. As labels of vagueness and approximation are present, fuzziness in the neighbour's closeness and the fuzzy rough set to form the FRNN. The advantages of using FRNN include insignificance of optimal value, better classification ability, non-dependence on prior structural details on the training data information, and better robust classification result.

2.5. Dataset description: The breast cancer (breast), diffuse large B cell lymphoma (DLBCL), small round blue-cell tumours (SRBCT), Leukemia datasets are used in the proposed system are extracted from the works [14]. Electroencephalography (EEG) and Gisette dataset are extracted from http://epileptologiebonn.de/cms/front_content.php?idcat=193&lang=3 and UCI machine learning repository also classified using the proposed system like cancer dataset. These datasets are used to measure the effectiveness of the classification model. The description of the dataset is shown in Table 1.

2.6. Evaluation parameters: The performance of the selected features is evaluated based on classification accuracy, sensitivity, specificity, precision and F-measure parameters.

- *Classification accuracy* is measured based on the proportion of number of correctly classified instances against the total number of instances.
- *Sensitivity* is gauged based on the proportion of number of correct positive predictions against the total number of positives, which is otherwise known as recall or true positive rate.

Table 1 Description of the datasets

	Breast	DLBCL	SRBCT	Leukaemia	EEG	Gisette
features	9217	4027	2309	12,583	4097	5000
instances	54	58	63	57	500	13,500
classes	5	6	4	3	5	2

- *Specificity* is looked at based on the proportion of number of correct negative predictions against the total number of negatives.
- *Precision* is arrived at based on the proportion of number of positive predictions against the total number of positives.
- *F-measure* is the harmonic mean of precision and recall.

3. Proposed model: The proposed classification model is conceived with the reduced feature is described in Fig. 1. The cancer datasets are collected from the repository and the data reduction is performed to select the contributing attribute for the classification [18, 19]. To support the functionality, a subset of features are selected by the proposed GSFR feature selection algorithm, which hybridises the property of evolutionary sequential genetic search technique and fuzzy rough set. Thereafter the reduced dataset model is generated by applying the FRNN classification. The model generated from the reduced features from the dataset foresees the cancer data with higher classification accuracy than the complete attributes without reduction [15].

Fuzzy rough feature selection identifies the appropriate set of features by eliminating the irrelevant features to improve the performance of the classifier [5]. The attribute subset is generated by the evolutionary sequential genetic search algorithm based on crossover and mutation probabilities. The steps involved in the generation and evaluation of subset are as follows:

- Randomly select the population (subset) from the dataset.
- Fitness function is constructed and evaluated to find the fitness of the population.
- Parent chromosome (features) is selected based on the precise value.
- Crossover probability of 0.80 is applied to crossover the parent chromosomes to generate the offspring (generate new subset).
- Mutation probability of 0.05 is to create the best population (subset).
- Fitness function is applied until it reaches the final shape.

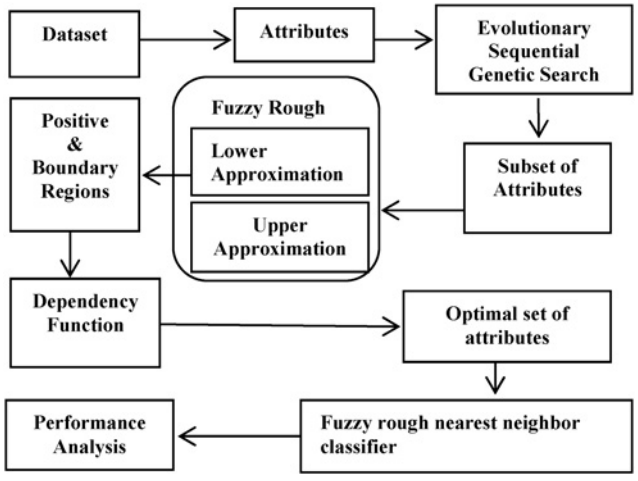

Fig. 1 *Block diagram denotes the classification model involving the feature selection by GSFR algorithm using genetic search technique and fuzzy rough set*

The initial subset for evaluation is randomly generated from the dataset and the generated subset is evaluated by the modified dependency function in the fuzzy rough set. The modified dependency function depends on the fuzzy rough degree of dependency by taking into account the positive region value and boundary region value (see Fig. 2).

The fuzzy rough dependencies of the attributes in (9) and (10) act as the fitness function in a genetic search algorithm to evaluate the subset of attribute. The evaluated subset is considered as the final one when it reaches the stopping criteria or else the crossover and mutation probabilities are applied to generate a new subset and evaluated with the fitness function. The same process goes on until it reaches the stopping criterion. FRNN classification is the combination of fuzzy rough approximation and fuzzy nearest neighbour (FNN) algorithm [5, 13]. If the lower approximation is high, almost all y objects neighbour belongs to the class. If the upper approximation is high, at least one y objects neighbour belongs to the class. The membership of test data objects is based on

Inputs: Dataset
Output: Selected Best Subset of features γ'_{best}

Begin
  Generation=0;
  Subset generated previously (γ'_{prev}) =0;
  Best subset (γ'_{best}) =0
  Randomly initialize the population (C)
    / initial subset of the dataset*/*
  $\gamma'_{prev} = \gamma'_{best}$
  Repeat
    Evaluate the subset by fitness function

$$\gamma_C(D) = \frac{\sum_{x \in U} \mu_{POS_C(D)}(x)}{U} + \mu_{BND_C(D)}(x)$$

    / evaluate the fitness of subset */*
    If (γ c(D) > γ'prev) then
      $\gamma'_{best} = \gamma_c(D)$
    EndIf
    If($\gamma'_{best} = = \gamma'_{prev}$) then
      */*stopping criteria*/*
      Subset generated goto step 22
    Else
      $\gamma'_{prev} = \gamma_c(D)$
      Select the parent from the population(C)
      Apply the crossover probability on parent
      Generate new population(C+1)
        / new subset is generated by crossover */*
      Do mutation probability on population(C+1)
        */*Mutation on the new population*/*
      Generation++;
      goto step 6
    EndIf
      / Fitness function is applied till it reaches the stopping condition*/*
  Until the best subset generated
  The best subset is γ'_{best}
End

Fig. 2 *Algorithm: proposed GSFR algorithm*

doi: 10.1049/htl.2018.5041

Inputs: Dataset and decision classes
Output: Classification of class(y)

1: A←Nearest Neighbor of y
2: $\tau \leftarrow 0$, Class←null
3: $\forall_X \in X$
4: $if\ (((\mu_{\overline{P}X}(F_i) + \mu_{\underline{P}X}(F_i))/2 \geq \tau)$
5: Class←X
6: $\tau \leftarrow ((\mu_{\overline{P}X}(F_i) + \mu_{\underline{P}X}(F_i))/2$
7: Output Class

Fig. 3 *Algorithm: FRNN algorithm*

the average of upper and lower approximations [16]. Then the average is compared with the higher existing value (τ). If the average value is higher, it needs to be replaced with an existing value. If it is otherwise, the process should continue (see Fig. 3).

The FRNN generate the model by including the vague quantifiers in the definition of upper and lower approximations. Then it is applied to non-exhausting 10-fold cross-validation where the complete dataset is involved where the model is tested in the training phase. In 10-fold cross-validation, the dataset is split into 10 equal sized samples i.e. $k=1, \ldots, 10$. Ten rounds of validation are performed in the first round, wherein one subset acts as test set and the remaining nine subsets are involved in model generation. In the next round, the subset involved in testing moves to training and one subset in training set involves in testing [10]. Similarly, remaining rounds of validation are performed, where all the subsets are involved in training and testing. The average of ten rounds of validation entails the final result.

4. Results and discussions: In order to analyse the performance of proposed GSFR feature selection algorithm the experiments are conducted using Weka tool with system configuration of Inte1 core i5-4590 processor, 3.00 GHz and 4 GB RAM.

4.1. Percentage of feature selection: In order to compare the number of features selected in the proposed GSFR feature selection algorithm with existing feature selection algorithms, experiments are conducted and results are presented in Table 2.

The number and percentage of features selected using various feature selection algorithms namely CFS [6], consistency subset evaluation-based feature selection (CSE) [5], principal component analysis-based feature selection (PCA) [8], Wrapper subset evaluation-based feature selection (WSE) [7] and the proposed GSFR feature selection algorithm for the different dataset are presented in Table 2. As the SRBCT and Leukaemia datasets have more redundant features, the feature selection techniques produce less features, when compared to other datasets. The difference in a number of features selected by the proposed GSFR feature selection algorithm and the average number of features selected by the existing algorithms, which read breast at 6%, DLBCL at 9%, SRBCT and Leukaemia at 4%, EEG at 13% and Gisette at 14%. This indicates that the proposed GSFR feature selection algorithm selects less number of features compared to other algorithms. This is due to the fact that the proposed GSFR feature selection algorithm deploys metaheuristic search property of genetic algorithm and applies modified dependency function in the fuzzy rough set.

4.2. Identification of suitable classifier: In order to identify the suitable classifier related experiments are conducted with proposed GSFR feature selection algorithm, without applying the proposed algorithm by employing various classifiers namely nearest neighbour (NN) [9], J48 [10], FNN [11], Random forest (RandF) [20], Gradient Booting (GradBoost) [21], SVM [12] and FRNN.

The outcome, in terms of classification accuracy and computation time, is presented in Table 3, wherein it is quite evident that in

Table 2 Datasets with reduced features considering different feature selection techniques

Dataset	Total number of features	Number of features selected with proposed and other feature selection techniques (% of features selected)				
		CFS [6]	CSE [5]	PCA [8]	WSE [7]	Proposed GSFR
breast	9217	2677 (29%)	1928 (21%)	1458 (16%)	1337 (15%)	1329 (14%)
DLBCL	4027	1014 (25%)	872 (22%)	543 (13%)	313 (8%)	306 (8%)
SRBCT	2309	254 (11%)	167 (7%)	96 (4%)	53 (2%)	47 (2%)
Leukaemia	12,583	1239 (10%)	767 (6%)	593 (5%)	193 (2%)	190 (2%)
EEG	4097	1075 (26%)	912 (22%)	527 (13%)	298 (7%)	155 (4%)
Gisette	5000	1202 (24%)	923 (18%)	519 (10%)	392 (8%)	65 (1%)

Table 3 Classification accuracy and computation time for different classifier with proposed GSFR feature selection algorithm A and without using proposed GSFR feature selection algorithm B

Dataset	Accuracy, % and time, s	NN [9]		J48 [10]		FNN [11]		RandF [20]		GradBoost [21]		SVM [12]		FRNN [13]	
		B	A	B	A	B	A	B	A	B	A	B	A	B	A
breast	acc.	60.37	66.67	70.37	73.19	66.67	70.37	62.96	72.22	73.78	73.90	73.54	74.04	74.02	74.07
	time	0.37	0.10	1.30	0.07	0.25	0.06	0.34	0.18	4.88	0.65	1.34	0.06	0.26	0.04
DLBCL	acc.	72.76	77.59	68.80	72.40	80.76	82.75	75.86	81.03	87.93	89.66	90.00	94.45	93.10	96.55
	time	0.12	0.08	0.44	0.06	0.12	0.04	0.09	0.01	2.44	0.18	0.12	0.05	0.11	0.02
SRBCT	acc.	66.19	74.60	72.54	74.60	67.30	76.19	72.54	75.71	75.71	78.65	78.89	79.41	79.15	80.95
	time	0.08	0.04	0.18	0.10	0.06	0.01	0.06	0.01	1.38	0.02	0.06	0.04	0.01	0.01
Leukaemia	acc.	67.72	73.68	61.19	65.78	56.74	57.90	73.68	75.44	70.70	81.23	72.48	82.98	70.17	87.71
	time	0.37	0.12	0.71	0.12	0.36	0.05	0.84	0.02	3.77	0.06	0.44	0.01	0.54	0.05
EEG	acc.	60.08	61.80	74.20	88.80	73.20	89.80	75.80	89.89	78.00	84.23	78.20	87.20	78.80	90.80
	time	1.18	0.05	1.28	0.36	1.75	0.05	1.41	0.12	1.17	0.18	1.42	0.39	1.82	0.05
Gisette	acc.	69.08	76.03	70.02	78.76	70.94	79.09	72.50	80.00	73.27	82.60	75.00	87.80	75.25	90.60
	time	1.34	0.12	1.46	0.19	1.12	0.09	0.90	0.05	1.02	0.06	0.50	0.04	0.24	0.01

Table 4 Performance analysis of FRNN with different feature selection techniques

Feature selection techniques	Dataset	Sensitivity	Specificity	Precision	F-measure	Accuracy
CFS [6]	Breast	0.653	0.889	0.709	0.660	66.67
	DLBCL	0.825	0.948	0.863	0.829	77.59
	SRBCT	0.709	0.884	0.711	0.709	71.43
	Leukaemia	0.735	0.848	0.721	0.726	73.68
	EEG	0.832	0.952	0.819	0.816	83.2
	Gisette	0.818	0.786	0.813	0.796	81.82
CSE [5]	Breast	0.687	0.892	0.675	0.666	68.52
	DLBCL	0.846	0.952	0.875	0.846	79.31
	SRBCT	0.732	0.892	0.728	0.726	73.02
	Leukaemia	0.747	0.861	0.743	0.743	75.44
	EEG	0.85	0.957	0.844	0.838	85
	Gisette	0.836	0.802	0.834	0.814	83.6
PCA [8]	Breast	0.685	0.905	0.793	0.706	70.37
	DLBCL	0.856	0.956	0.884	0.859	81.03
	SRBCT	0.747	0.901	0.749	0.744	74.60
	Leukaemia	0.761	0.872	0.761	0.758	77.19
	EEG	0.866	0.962	0.865	0.854	86.6
	Gisette	0.843	0.822	0.847	0.830	84.26
WSE [7]	Breast	0.739	0.912	0.796	0.728	72.22
	DLBCL	0.878	0.96	0.877	0.873	82.76
	SRBCT	0.767	0.907	0.788	0.760	76.19
	Leukaemia	0.856	0.923	0.864	0.858	77.78
	EEG	0.882	0.967	0.877	0.869	88.2
	Gisette	0.871	0.846	0.878	0.857	87
proposed GSFR	Breast	0.741	0.948	0.826	0.781	74.07
	DLBCL	0.942	0.985	0.904	0.923	96.55
	SRBCT	0.849	0.94	0.841	0.845	80.95
	Leukaemia	0.705	0.856	0.746	0.725	87.71
	EEG	0.908	0.977	0.916	0.908	90.8
	Gisette	0.906	0.85	0.91	0.904	90.6

FRNN classifier, the classification accuracy is improved upon with lesser computation time (A). For instance, the difference between the classification accuracy of FRNN classifier and the average of other existing classifier for DLBCL dataset works out to 13.57% and the computation time differs by 0.05 s. Similarly, it is apparent that the FRNN produces improved classification accuracy with lesser computation time against other datasets. This is due to the fact that the modified dependency function of the fuzzy rough set is adopted as a fitness function in the proposed GSFR feature selection algorithm.

Since the FRNN classifier outperforms all other classifiers, the performance analysis in terms of sensitivity, specificity, precision, recall, F-measure and accuracy for the existing feature selection algorithm and proposed GSFR feature selection algorithm is analysed thereon by using FRNN classifier. The results are shown in Table 4.

4.3. Performance comparison: The selected features from the proposed GSFR feature selection algorithm achieve improved classification accuracy with lesser computation time which aligns with FRNN classifier. Thereafter, it is applied to measure the performance of the existing and proposed feature selection algorithm as shown in Table 4. To support this claim, the difference between classification accuracy of the proposed GSFR feature selection algorithm and the average of other existing feature selection algorithm for the SRBCT dataset produces the improved accuracy of 7.14%. It implies that the proposed GSFR feature selection algorithm selects the contributing features for classification.

5. Conclusion: This Letter prescribes the cancer classification model developed from FRNN classifier using the reduced feature from proposed GSFR feature selection algorithm based on the hybridisation of evolutionary sequential genetic search algorithm and fuzzy rough set. The process aims to reduce the number of features from the datasets and thereafter a model is developed and validated by 10-fold cross-validation technique. The proposed GSFR feature selection algorithm and classification model offers proven results compared to other techniques in terms of a number of features, classification accuracy, precision, recall, F-measure and computation time. As it applies lesser features, the time consumed to develop the proposed model is minimal. The proposed GSFR feature selection algorithm selects ideal features which result in improved accuracy with lesser computational time. This Letter opens new areas of further study with specific reference to hybridise fuzzy rough set with both particle swarm optimisation and ant colony optimisation to improve upon still further.

6. Funding and declaration of interests: None declared.

7 References

[1] Salem H., Attiya G., El-Fishawy N.: 'Classification of human cancer diseases by gene expression profiles', *Appl. Soft Comput.*, 2017, **50**, pp. 124–134

[2] Hosseini S., Turhan B., Mantyla M.: 'A benchmark study on the effectiveness of search-based data selection and feature selection for cross project defect prediction', *Inf. Softw. Technol.*, 2018, **95**, pp. 296–312

[3] Herrera-Semenets V., Pérez-García O.A., Hernández-León R., *ET AL.*: 'A data reduction strategy and its application on scan and backscatter detection using rule-based classifiers', *Expert Syst. Appl.*, 2018, **95**, pp. 272–279

[4] Karanasiou G.S., Tripoliti E.E., Papadopoulos T.G., *ET AL.*: 'Predicting adherence of patients with HF through machine learning techniques', *Healthc. Technol. Lett.*, 2016, **3**, (3), pp. 165–170, doi: 10.1049/htl.2016.0041

[5] Onan A.: 'A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer', *Expert Syst. Appl.*, 2015, **42**, pp. 6844–6852
[6] Jain I., Jain V.K., Jain R.: 'Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification', *Appl. Soft Comput.*, 2018, **62**, pp. 203–215
[7] Korfiatis V.Ch., Asvestas P.A., Delibasis K.K., *ET AL.*: 'A classification system based on a new wrapper feature selection algorithm for the diagnosis of primary and secondary polycythemia', *Comput. Biol. Med.*, 2013, **43**, pp. 2118–2126
[8] Narendar Reddy K., Ravi V.: 'Differential evolution trained kernel principal component WNN and kernel binary quantile regression: application to banking', *Knowl.-Based Syst.*, 2013, **39**, pp. 45–56
[9] Chuang L.-Y., Yang C.-S., Wu K.-C., *ET AL.*: 'Gene selection and classification using Taguchi chaotic binary particle swarm optimization', *Expert Syst. Appl.*, 2011, **38**, pp. 13367–13377
[10] Baitharua T.R., Panib S.K.: 'Analysis of data mining techniques for healthcare decision support system using liver disorder dataset'. *Procedia Computer Science*, 2016, **85**, pp. 862–870
[11] Kumar M., Rath S.K.: 'Microarray data classification using fuzzy K-nearest neighbor'. IEEE Int. Conf. on Contemporary Computing and Informatics (IC3I), Mysore, India, November 2014
[12] Zhou W., Dickerson J.A.: 'A novel class dependent feature selection method for cancer biomarker discovery', *Comput. Biol. Med.*, 2014, **47**, pp. 66–75
[13] Qu Y., Shen Q., Parthalain N.M., *ET AL.*: 'Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels', *Int. J. Approx. Reason.*, 2013, **54**, pp. 184–195
[14] Wei J.-M., Wang S.-Q., Yuan X.-J.: 'Ensemble rough hypercuboid approach for classifying cancers', *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, pp. 381–391
[15] Maa B., Xia Y.: 'A tribe competition-based genetic algorithm for feature selection in pattern classification', *Appl. Soft Comput.*, 2017, **58**, pp. 328–338
[16] Jensen R., Cornelis C.: 'Fuzzy-rough nearest neighbour classification and prediction', *Theor. Comput. Sci.*, 2011, **412**, pp. 5871–5884
[17] Dai J., Xu Q.: 'Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification', *Appl. Soft Comput.*, 2013, **13**, pp. 211–221
[18] Palese L.L.: 'A random version of principal component analysis in data clustering', *Comput. Biol. Chem.*, 2018, **73**, pp. 57–64
[19] Sau A., Bhakta I.: 'Predicting anxiety and depression in elderly patients using machine learning technology', *Healthc. Technol. Lett.*, 2017, **4**, (6), pp. 238–243, doi: 10.1049/htl.2016.0096
[20] Wade B.S.C., Joshi S.H., Gutman B.A., *ET AL.*: 'Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods', *Pattern Recognit.*, 2017, **63**, pp. 731–739
[21] Liu Y., Gu Y., Nguyen J.C., *ET AL.*: 'Symptom severity classification with gradient tree boosting', *J. Biomed. Inf.*, 2011, **75**, pp. S105–S111