ACKNOWLEDGEMENT

First and foremost, I wish to express my deepest gratitude to my previous supervisor Dr.Siew Chin Neoh, for her endless support during this research study. She has spent her valuable time and efforts to train my research skills, and provided constructive and challenging feedbacks to improve my research work. This thesis would not have been possible at all without her constant encouragement and guidance.

I would like to thank my husband whose continuous support for all these years has been well beyond the call of duty. His constant support, patience, and advice have moulded me to finish this research with ease. He has been my source of encouragement throughout the course of study. I would like to express my sincere gratitude to him for helping me improve my research writing skills.

I wish to extend my thanks to my present supervisor Dr.Asral B. Bahari Jambek, for his prompt help and review of this thesis whenever I encounter problems. His continuous support, advices, and patience have been of utmost importance for finishing my thesis.

Finally, I dedicate this thesis to my husband, parents and my daughter.

I sincerely acknowledge the FRGS fund (9003-00261) received from the Ministry of Higher Education, Malaysia, through UniMAP.

TABLE OF CONTENTS

CHAPTER 3 RESEARCH METHODOLOGY

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BBFS	Branch and Bound Feature Selection
BFS	Backward Feature Selection
BN	Bayesian Network
BT	Breast Tissue
CAD	Cleveland Heart Disease
CTG	Cardiotocogram
ELM	Extreme Learning Machine
ES	Erythemato - Squamous
FHR	Fetal heart rate
FFS	Forward Feature Selection
FR	Feature Reduction
FS	Feature Selection
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
IFS	Individual Feature Selection
IGA	Improved Genetic Algorithm
k-NN	*k*-Nearest Neighbor
LDA	Linear Discriminant Analysis
MEEI	Massachusetts Eye and Ear Infirmary
PCA	Principle Component Analysis
PD	Parkinson’s disease
RBF	Radial Basis Function
AUC	Area Under Receiver Operating Characteristic Curve

SBS	Sequential Backward Selection
SD	Standard Deviation
SFS	Sequential Forward Selection
SVM	Support Vector Machine
SUS	Stochastic Universal Sampling
UC	Uterine contractions
UCI	University of California, Irvine

Penyiasatan Mengenai Evolusi Penambahbaikkan Teknik Pemilihan Ciri-ciri untuk Aplikasi Bioperubatan

ABSTRAK

Dalam hipotesis pengecaman corak , pengelasan set data perubatan menjadi satu tugas yang mencabar disebabkan bilangan ciri-ciri yang banyak dan batasan dalam latihan data. Pertindihan yang terdapat dalam ciri-ciri yang tidak relevan ini menjejaskan ketepatan pengelasan secara keseluruhan. Isu-isu yang tidak dapat diatasi dalam klasifikasi boleh diselesaikan dengan teknik Pemilihan Ciri-ciri (FS) yang dilaksanakan melalui kaedah pengoptimuman kombinatorik seperti Algoritma Genetik (GA). Kaedah FS yang cekap dapat menyingkirkan data yang tidak relevan dan tidak diperlukan daripada set data, mengurangkan bilangan ciri-ciri, dan menghasilkan ketepatan pengelasan yang boleh diterima. Walau bagaimanapun ,terdapat banyak batasan dalam GA Asas seperti penumpuan pramatang , darjah ketepatan penyelesaian yang rendah disebabkan lintas yang tetap dan kadar mutasi, kapasiti penumpuan yang rendah, kekurangan kepelbagaian populasi dan sebagainya. Batasan ini menyekat pencarian genetik dari mencapai penyelesaian optimum global dan dengan itu mengurangkan kecekapanya sebagai kaedah FS. Untuk mengatasi masalah ini, tesis ini mencadangkan beberapa penyelesaian yang boleh diterima untuk menyelesaikan isu-isu ini iaitu pertamanya, dengan menambah teknik skala kecergasan (dipanggil penskalan sigma) bersama-sama dengan fungsi pemilihan SUS (Persampelan Stochastic Universal). Teknik berskala ini telah membolehkan pencarian genetic tersebut menyesuaikan semula nilai-nilai kecergasan populasi, yang telah menghalang GA dari mencapai penumpuan pramatang. Kedua, darjah ketepatan penyelesaian bertambah baik dengan mengubah-suai lintas dan kadar mutasi berdasarkan kecergasan populasi. Untuk mengukur kualiti kromosom, tiga fungsi kecergasan yang berbeza iaitu ketepatan Klasifikasi, min geometri dan pengagregatan wajaran min Geometri dan penjumlahan ciri-ciri terpilih telah digunakan. Penjodoh bekerja bersama-sama dengan sepuluh kali ganda kaedah pengesahan silang telah berkhidmat sebagai penilai algoritma IGA (Algoritma Genetik ditambahbaik) yang dicadangkan. Keputusan eksperimen dibentangkan dari segi beberapa langkah prestasi seperti ramalan positif, ramalan negatif, Kepekaan, Spesifikasi, Ketepatan, F-mengukur, AUC, statistik Kappa dan G-min. Melalui kaedah IGA yang dicadangkan, ketepatan pengelasan yang memberangsangkan telah diperolehi untuk kesemua enam dataset penanda aras iaitu 100% untuk Meei dataset, 99.49% bagi dataset PD, 84%, untuk CAD dataset, 99.24% bagi dataset ES, 94,34% bagi dataset BT & 94% untuk dataset CTG. Selain itu, pengurangan ciri subset juga telah dicapai untuk kesemua dataset iaitu 8 ciri-ciri untuk dataset Meei & PD, 3 daripada 9 ciri-ciri untuk dataset CAD, 14 daripada 34 ciri-ciri untuk dataset ES, 3 daripada 13 ciri-ciri untuk dataset BT dan 6 daripada 22 ciri-ciri untuk dataset CTG masing-masing. Hasil kajian ini menunjukkan peningkatan prestasi pengkelasan yang ketara dan kaedah yang dicadangkan melangkaui prestasi kerja-kerja yang pernah diterbitkan sebelum. Semua algoritma yang digunakan dalam eksperimen ini telah disimulasikan menggunakan MATLAB 2011a.

Investigation of Improved Evolutionary Feature Selection Techniques for Biomedical Applications

ABSTRACT

In the hypothesis of pattern recognition, the classification of medical datasets is becoming a challenging task due to the large number of features and training data limitations. Redundancies present in these irrelevant features affect the overall classification accuracy. Such insurmountable issues to classification can be overcome by Feature Selection (FS) techniques performed through combinatorial optimization methods such as Genetic Algorithm (GA). However, there are many limitations in basic GA like premature convergence, low degree of solution accuracy due to fixed crossover and mutation rates, low convergence capacity and lack of population diversity. These limitations restrict the genetic search from reaching global optimal solution and thereby reducing their efficiency as a FS method. To overcome these issues, this thesis suggests an effective wrapper framework with some acceptable solutions i.e. firstly, by adding fitness scaling technique (called sigma scaling) along with the Stochastic Universal Sampling (SUS) selection function. This scaling technique has enabled the genetic search to re-adjust the fitness values of the population, which has prevented GA from reaching premature convergence. Secondly, the degree of solution accuracy is improved by adaptively changing the crossover and mutation rates based on the population fitness. Further, the masking concepts of crossover and mutation enhanced the population diversity and increased the convergence capacity as well. To measure the quality of the chromosomes, three different objective functions namely Classification Accuracy, Geometric Mean and a weighted aggregation of Geometric mean and sum of selected features have been employed. The classifiers employed along with ten-fold cross validation method have served as an evaluator of the proposed Improved Genetic Algorithm (IGA) algorithm. The experimental results are presented in terms of several performance measures like Positive prediction, Negative prediction, Sensitivity, Specificity, Accuracy, F-measure, AUC, Kappa statistic and G-mean. Through the proposed IGA method, promising classification accuracy has been obtained for all the six benchmark datasets i.e. 100% for MEEI dataset, 99.49% for PD dataset, 84% for CAD dataset, 99.24% for ES dataset, 94.34% for BT dataset & 94% for CTG dataset. Also, a reduced feature subset has been attained for all these datasets as well i.e. 8 features for MEEI & PD dataset, 3 out of 9 features for CAD dataset, 14 out of 34 features for ES dataset, 3 out of 13 features for BT dataset and 6 out of 22 features for CTG dataset has been obtained respectively. On the whole, experimental results show that the proposed algorithm has evolved a feature subset with a smaller number of features and higher classification performance than using all the features. All the algorithms used in these experiments were simulated using MATLAB 2011a.

CHAPTER 1

INTRODUCTION

This chapter introduces a new search algorithm developed for classifying the binary and multiclass problems using genetic based feature selection techniques. It tells about the purpose of Genetic Algorithm (GA) in solving pattern recognition problems and their significance in this specific problem domain. In addition, this chapter summarizes the objectives of the proposed research and organization of the thesis.

1.1 PREFACE

Data dimensionality is one of the core problems, which usually occurs when large datasets are used for solving pattern recognition applications. This happens due to the infeasibility of populating the feature space (sufficiently) with limited data. To overcome this issue, many real-time practices have adopted a technique called data reduction, which means reducing the innumerable amounts of data into meaningful parts. This can be achieved by either Feature Transformation (transforming the features to populate a new, lower dimensional space), or by Feature Selection (FS).

In this research work, we are going to have a brief outlook on feature selection in the following discussion: FS is the process of removal of irrelevant, redundant and noisy features present in the main features of data. The reason behind such removal is mainly due to the fact that those redundant features will not achieve the best recognition rate and further lead to the difficulty in data interpretation (Lee, Jeong, & Hahn, 2007).

For this aim, reduction of data demands searching an optimal (the most feasible solution within the feasible region) subset of features and it is highly advantageous in many aspects like:

(a) Dimension reduction of the feature subset to reduce the computational cost;

(b) Noise reduction to improve the prediction accuracy;

(c) Finding more interpretable features or characteristics which easily recognize the target description;

(d) To render faster and more cost-effective models and

(e) To gain a deeper understanding into the underlying processes that has generated the data.

1.2 PROBLEM STATEMENT

In pattern recognition tasks, classification accuracy is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. Classification problems have a large number of features among which not all of them are useful for classification. Also, the irrelevant and redundant features have the possibilities of reducing the classification accuracy and lessening the quality of the whole feature set. To overcome these issues, Feature selection is proposed to select the highly discriminative features, which in-turn increase the quality of the feature space and classification performance. By selecting those relevant features, FS techniques decrease the data dimensionality and shorten the running time as well.

However, at certain instance, FS becomes difficult to carryout due to feature interaction, high computational cost due to large search space and stagnation in local optima. Therefore, an efficient global search strategy is required to ease the task of FS. Nowadays, evolutionary optimization techniques are a group of powerful global search algorithms, which have been successfully employed to solve many pattern recognition tasks. Hence, in this regard, GA has been applied for solving such FS problems.

1.3 MOTIVATION OF THE STUDY

There are plenty of FS techniques used to resolve many pattern recognition applications and real time problems of artificial intelligence (Papakostas, Koulouriotis, Polydoros, & Tourassis, 2011). The selection of discriminative features is highly decisive to improve the recognition rate in such problems. A good FS method will perform this selection of features in an efficient way. In this regard, Genetic algorithm (GA) is gaining more importance in optimizing the features subset within the pool of feature vectors.

The reason of choosing GA as a FS method is due to its extreme capability in handling the global search problems in vast and complex spaces. In fact, in order to handle irregular and complex search spaces, the search must adopt a global strategy and rely heavily on intelligent randomization and GAs follow such a strategy. In GA, the search procedure is biased towards the global optimal solution which makes it highly robust within a multimodal landscape. However, being a very useful technique in feature selection of many aspects, still, there is a room for GA to prove its implication in certain genres of pattern recognition applications. To narrow down this, we have attempted to perform the pattern classification of binary and multiclass problem on various benchmark and medical datasets in this research work.

1.4 OBJECTIVES OF THE STUDY

The main objectives that have to be carried out through the proposed research work are:

(a) To perform the pattern classification of medical datasets through an expeditious optimization algorithm and thereby to achieve a promising classification accuracy.

(b) To choose the optimal feature subset from the existing features which best describes the target concept.

1.5 SCOPE OF THE STUDY

The major scope of this study is to investigate the significance of GA in feature selection for solving the binary and multiclass problems. This research work is an effort

(a) To develop an efficient search algorithm through Improved GA (IGA), to solve the classification of medical datasets.

(b) To assess the strength of the IGA algorithm by experimenting it on distinct datasets and to gauge its efficiency in terms of performance measures like sensitivity, specificity, accuracy, F-measure, Geometric Mean (G-Mean), Kappa statistic, AUC (Area Under Receiver Operating Characteristic curve).

1.6 THESIS ORGANISATION

This thesis investigates the IGA based pattern classification techniques. The research conducted is outlined and presented in five chapters of this thesis. The rest of the chapters are outlined as the following:

Chapter 2 describes the basic GA process and its previous studies on pattern classification. It exemplifies the significance of IGA and its modified objective function used in this proposed work. In addition, the explanation on the datasets in this research work and the experimentation of the proposed algorithm on various datasets are given.

Chapter 3 reports about the proposed IGA method in the classification of pattern recognition tasks. The classifiers used in the performance evaluation and the implementation of validation methods are described. A block diagram that depicts the entire process of the suggested work is given together with elaboration and justification on the usage of different types of datasets.

Chapter 4 elaborates the experimental results of the proposed IGA method. It gives a comparative study of both basic GA and IGA in various aspects. In addition, it outlines the validation methods conducted and summarizes the best performance measures achieved through the simulation results.

Chapter 5 concludes the proposed work and suggests the future works in the perspective to enlarge the application areas and to enhance the performance of the proposed algorithm.

CHAPTER 2

LITERATURE REVIEW

This chapter explains the previous studies carried out by Genetic Algorithm (GA) and their application towards the suggested work in the classification of binary and multiclass problems. It describes the contribution of Adaptive or Improved GA (IGA) in this specific subject. Also, this chapter outlines the significance of IGA and the observation from the previous works.

2.1 INTRODUCTION

GAs has been used in a wide range of engineering and scientific fields to quickly provide usable solutions in the combinatorial optimization problems. In this work, an efficient search algorithm named as IGA has been proposed, developed and implemented for a particular pattern classification issue and its importance as a FS method has been elaborated in detail.

2.2 BASIC CONCEPTS OF GENETIC ALGORITHM

Basic GA deals with candidate solutions which are represented by individuals (or chromosomes) in a large population. It starts with the random generation of initial set of chromosomes followed by their corresponding fitness evaluation. The successive generations are created (iteratively) by the highly fit individuals of the current generation. On achieving the eligible individual that satisfies our constraint, the GA cycle is stopped using the termination criteria and this individual becomes the solution of the problem.

Selection, crossover and mutation are the significant genetic operators which help in the reproduction process of the chromosomes (Sivanandam & Deepa, 2007).

Ideally, a GA cycle takes place using these genetic operators and such genetic parameters influencing this process tend to produce good individuals. A Basic GA cycle is depicted in Fig.2.1.

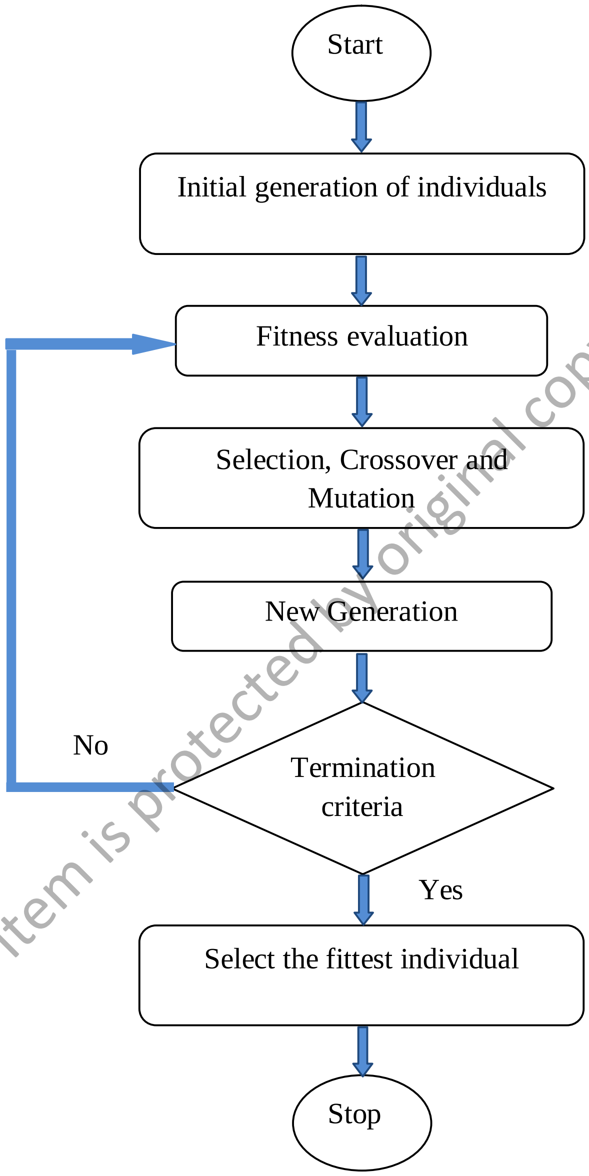
Start
Initial generation of individuals
Fitness evaluation
Selection, Crossover and Mutation
New Generation
No
Termination criteria
Yes
Select the fittest individual
Stop

Figure 2.1: Basic GA cycle

In classification problems, a chromosome is generally encoded as binary strings: sequences of 1’s and 0’s. If ‘1’ appears, it means that the corresponding feature is selected and if ‘0’ appears, it means that the corresponding feature is rejected. The fittest individuals get the higher chance of being produced in further generations. Some of the GA operators and the respective parameters that determine the genetic flow of the algorithm are discussed in the following:

2.2.1 Selection (or Reproduction) process chooses the parent chromosomes for the next generation based on their fitness value. Once after the first generation of chromosomes is initialized, the selection process takes place based upon the fitness values of the chromosomes. As a result, the fittest ones are forwarded to the next successive generations (Srinivasan, Ramalingam, & Sellam, 2012).

2.2.2 Crossover (or Recombination) combines the selected individuals to generate better offspring by exchanging the genetic information between them, i.e. the parent chromosomes exchange their genetic material, resulting in the production of child chromosomes. This occurs with a crossover probability P_c, when the parent chromosomes are chosen for breeding (Xie & Wang, 2011).

2.2.3 Mutation swaps the chromosomal bits as 0s into 1s and 1s into 0s, with a mutation probability, P_m. It restores the lost or unexplored genetic material of the population and hence maintains the genetic diversity. It assists GAs to escape from the local optima and to proceed further in the search towards the global optima (Kohavi & John, 1997).

2.2.4 Termination criteria (or Stopping criteria) determines the flow of GA cycle i.e. the crucial points, where the algorithm has to stop and resume accordingly (Karegowda, Manjunath, & Jayaram, 2010). Some of the termination criteria which stops this evolutionary process are

(a) On reaching specified number of generations;

(b) On achieving an anticipated classification accuracy;

(c) On obtaining a good and an optimal feature subset for the solution;

(d) When the change (addition or deletion of features) of feature subsets does not produce a better subset further.

2.2.5 Elitism is the concept of maintaining fittest individuals of every generation throughout the entire GA process using elitist chromosomes. These elites will be hired consequently in order to give birth to the children of the next generation. By using elitist strategy, the best individual in each generation is ensured to be passed to the next generation.

The target of the entire GA cycle lies in finding the highly fittest chromosome which either minimizes or maximizes the objective function as per the requirement of the problem domain. As fittest individuals are preserved and mated with one another, the average performance of individuals in the population is expected to increase.

2.3 PREVIOUS RESEARCH WORKS ABOUT GA

In classification tasks, discriminative features are extremely essential in describing the target concept and such a feature subset becomes instrumental in attaining the higher classification accuracy. GA helps in reaching this accuracy through its extreme search ability in the wide solution space. Generally, high dimensional features are not suitable for the task of classification problems.

In such cases, FS is necessary to reduce the data dimensionality which in turn achieves a promising recognition rate and this was investigated by Casale et al. (Sanchez-Monedero et al., 2010) in 2007. In his work, he has discriminated the emotional or stressed state of a speaker into different states like neutral, angry and loud and Lombard using Speech under Simulated and Actual Stress database. The inverse of the separation index (J^{-1}) has been used as the objective function for fitness evaluation of the speech features. On the other hand, Tiwari and Singh (2010) have explained the utility of GA in the field of correlation based feature selection(Sanchez-Monedero, Gutierrez, Fernandez-Navarro, & Hervas-Martinez, 2011). In the detection of vocal fold pathology, GA has been used to analyze the ability of acoustic parameters using various feature reduction and feature classification methods (Srinivasan et al., 2012). The Gaussian kernel of the Support Vector Machine (SVM) classifier has classified the optimized features into pathological and normal with accuracy of 83.3%.

Also, wavelet transformation has been assistive in guiding GA to detect the vocal fold pathology (Casale, Russo, & Serrano, 2007). In this investigation, feature extraction is carried out using MFCC (Mel-frequency Cepstral coefficient) and WPD (Wavelet Packet Decomposition). The error rate of the GMM (Gaussian Mixture Model) classifier is taken as the fitness function in GA and this genetic based feature selection has given a better accuracy of 91.54% through GMM's classification when compared to PCA (Principle Component Analysis). As an extension of this work, GMM classifier has been replaced by SVM classifier. This has achieved a better classification accuracy of 99.23% with a vector length of 22 from the original feature count of 139 features (Ocak & Ertunc, 2013).

In (Jashki, Makki, Bagheri, & Ghorbani, 2009), the researchers have accomplished the genetic framework in the fusion of multiple feature selection for the classification purpose. In this work, experiments were performed on three data sets (Colon cancer dataset, Prostrate cancer dataset and Ionosphere dataset) using three existing FS methods (Entropy based ranking, T-statistics and SVM-RFE). The weighted fitness function has been used in this algorithm with the aim of maximizing the classification accuracy and minimizing vector size i.e. number of features selected.

Apart from solving two-class problems, GA has been applied in the discrimination of multiclass problems as well. For instance, the categorization of voice signals into four classes like normal, unilateral vocal fold paralysis, vocal fold polyp and vocal fold nodules has been presented in (Khadivi Heris, Seyed Aghazadeh, & Nikkhah-Bahrami, 2009). In this study, the feature vectors are selected based on the Davies-Bouldin index (DBI) and fed to the SVM and *k*-NN classifiers, which gives an accuracy of 91% through SVM classification.

GA is highly assistive in solving the time complexity issues of the classification tasks. This concept has been analyzed by Li. S (Legg, Hutter, & Kumar, 2004), in which two datasets like Indian pine dataset and Washington DC mall dataset are used for this purpose. Firstly, conditional mutual information is applied to segment the bands into disjoint subspace, thus minimizing the search space. Secondly, the combined scheme of GA and SVM has assisted to search for the optimal combination of bands. Thirdly, the branch and bound search algorithm (BB) is used to prune the irrelevant bands and the resulting minimal set of relevant bands are classified and achieved an accuracy of above 90%.

Zhuo et al. (2008) has inferred that a combined scheme of GA and SVM can be applied for feature selection of the hyper spectral data. The GA-SVM method was realized using the ENVI/IDL language, and was then tested by applying to a HYPERION hyper spectral image. The number of bands used for categorization has been cut down from 198 to 13, while the classification accuracy has increased from 88.81% to 92.51% (Georgoulas, Stylios, Nokas, & Groumpos, 2004). Table 2.1 displays the various specifications used (by GA) in the previous works of (genetic based) pattern classification and Table A.1 shows the different fitness functions used by GA for classification problems and it has been described in Appendix A. This table presents the various kinds of fitness functions employed by the researchers for the purpose of genetic optimization. There are many genetic factors within a GA procedure, which determine the speed and efficacy of the algorithm. Thus, besides being a very useful technique in FS of pattern recognition and artificial intelligence applications, still, there is a room for GA to prove its implication in certain disciplines.

Table 2.1: Genetic specifications of Basic GA in previous works.

Methods	Crossover rate	Mutation rate	Crossover type	Mutation type	Population size	Runs	Selection method
Iterative filter-wrapper approach	1 .0	0.001	Single point	Random	20-30	10	Roulette wheel
Hidden Markonikov Model approach	0.98	0.02	Random	-	50, 100	40	Tournament
Fitness Uniform Selection Scheme	0.8	0.2	Single point	Random	30	100	-
Wavelet packet Decomposition	-	-	Scattered	Gaussian	1500	300	Stochastic uniform
Wavelet packet tree approach and LDB	0.2	-	Single point	Gaussian	300	30	-
GA based Wavelet packet approach	0.2	0.7	Scattered	Gaussian	100	100	-

From the literature review, it is observed that certain FS methods are claimed to be inefficient because the algorithm does not take feature dependency into account. In addition, the computational complexity is gradually increased along with the number of