Inteligencia Artificial 23(65), 100-114
doi: 10.4114/intartif.vol23iss65pp100-114

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Ensemble Feature Selection for Breast Cancer Classification using Microarray Data

Supoj Hengpraprohm[1,A] and Suwimol Jungjit[2,B]

[1]Data Science Program, Faculty of Science and Technology,
Nakhon Pathom Rajabhat University, Nakhon Pathom, Thailand
[2]Department of Computer and Information Technology, Faculty of Science,
Thaksin University, Phatthalung, Thailand
[A]supojn@webmail.npru.ac.th, [B]suwimol@tsu.ac.th

Abstract This paper proposes an ensemble filter feature selection approach, EnSNR, for breast cancer data classification. The Microarray dataset used in the experiments contains 50,739 features (genes) for each of 32 patients. The main idea of the EnSNR approach is to combine informative features which are obtained using two different sets of feature evaluation criteria. Features in the EnSNR subset are those features which are present in both sets of evaluation results. Entropy and SNR evaluation functions are used to generate the EnSNR feature subset. Entropy is a measure of the amount of uncertainty in the outcome of a random experiment, while SNR is an effective function for measuring feature discriminative power. Entropy and SNR functions provide some advantages for the EnSNR approach. For example, the number of features in the EnSNR subset is not user-defined (the EnSNR subset is generated automatically); and the operation of the EnSNR function is independent of the type of classification algorithm employed. Also, only a small amount of processing time is required to generate the EnSNR feature subset. A Genetic Algorithm (GA) generates the breast cancer classification 'model' using the EnSNR feature subset. The efficiency of the 'model' is validated using 10-Fold Cross-Validation re-sampling. When the 'EnSNR' feature subset is used, as well as giving a high degree of prediction accuracy (the average prediction accuracy obtained in the experiments in this paper is 86.92 ± 5.47), the EnSNR approach significantly reduces the number of irrelevant features (genes) to be analyzed for cancer classification.

Keywords: Ensemble approach, Feature selection, Microarray data, Genetic Algorithm, Cancer Classification.

1 Introduction

Breast cancer is the most common cancer in women. The reason for carrying out the research described in this paper is to improve on the data classification prediction performance so far achieved [1, 2]. This paper demonstrates that the proposed 'Ensemble' feature selection approach, 'EnSNR', is superior to the traditional 'Entropy' or 'Signal to Noise Ratio (SNR)' approaches, for the selection of informative features to be used in the prediction process. The feature selection and data classification system block diagram for the experiments is shown in Figure 1.

The block diagram shows:

- Breast Cancer Microarray Dataset. This is the source of patient data used in the experiments
- Feature Selection functions 'Entropy' and 'Signal to Noise Ratio (SNR)'
- Feature Selection process 'Ensemble (EnSNR)'

ISSN: 1137-3601 (print), 1988-3064 (on-line)

- A Genetic Algorithm (GA). This is the GA-Based Classification Algorithm which performs classification and validation tasks
- 10-Fold Cross-Validation

In this paper, the names 'features' and 'genes' mean the same thing.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of the Microarray and Microarray dataset. Section 3 explains feature selection. The genetic algorithm is described in Section 4. Section 5 deals with the 10-Fold Cross-Validation procedure. The experimental set-up is described in Sections 6. Results are discussed in Section 7. Conclusions and further action are given in Section 8.

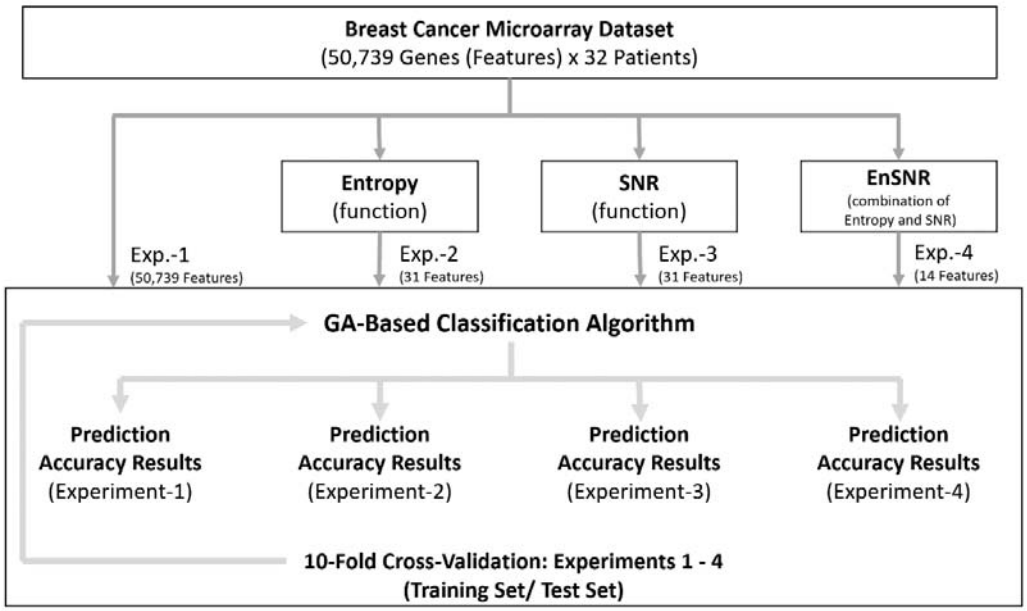

Figure 1. Feature Selection and Data Classification System Block Diagram.

2 The Microarray and Microarray Dataset

Gene expression levels can help a physician to diagnose a patient's condition [3, 4]. Microarray is a chip-based technology which is used to study the expression levels of genes: tens of thousands of genes are measured simultaneously. The experiments in this work use the Microarray dataset proposed by Junjie Fu et al. [5], as the input to the GA-Based Classification Algorithm proposed by Hengpraprohm [1]. The dataset, which is available thanks to Human Gene Expression Microarray technology, is available on the web site given in Junjie Fu's paper. Breast tissue samples obtained from 32 female patients are present in the Microarray. There are 50,739 genes per patient.

The microarray heat map is arranged in a matrix of zones. There are 32 columns in the matrix. A column is assigned to each patient. There are four column groups. Three groups are assigned to patients having different breast cancer subtypes. The subtypes are Luminal A, Luminal B and Triple negative. The fourth group is for patients with no breast cancer present in their tissues. There are 50,739 rows. Each row is assigned to a gene in the patient tissue.

On the heat map, there are 50,739 gene zones in each column. In a zone, the expression level of each gene is given by light colour/ colour intensity. Zone colour (red, green or black)/ colour intensity (applicable to red and green zones, only) represents the expression level. The dataset contains a positive expression level for a gene according to the light intensity of a red zone (the greater the intensity, the higher the positive value), and a negative expression level according to the light intensity of a green zone (the greater the intensity, the lower the negative value). Black is a 'zero' mean value.

In the dataset available on the web site, each patient is identified by a 'Patient ID'. The 'Patient Cancer Status' consisting of cancer subtype (see above), or 'no-cancer', is also given for each patient. Genes are identified by 'Gene ID' and 'Gene Symbol'. Gene expression level values are arranged in a 32 column (one column per patient) by 50,739 rows (one row per gene) matrix. Column data provides patient gene profiles consisting of 50,739 gene

expression levels per patient. In the dataset, expression level data given for a gene is a positive numerical value (i.e. over expression level), a negative numerical value (i.e. under expression level), or 'zero' (0) which is a mean gene expression level; in accordance with the gene's colour/ colour intensity on the microarray heat map.

3 Feature Selection

Feature selection is a crucial process which employs an evaluation criterion to choose a relevant feature subset with the aim of reducing the possibility of classification model overfitting. In addition, feature selection tends to improve the prediction performance of the classification model by removing irrelevant features to reduce computational time and increase classification accuracy [6-12]. Researchers use feature selection to improve the performance of classification algorithms [13, 14]. There are three approaches to feature selection; (1) the filter approach, (2) the wrapper approach, and (3) the embedded approach [6-8]. The most straightforward one is the "filter" approach. An advantage of this approach is that it evaluates the candidate features (genes) or candidate feature subsets (groups of genes) independently of the algorithm classifiers.

Recently, Genetic Algorithms have been adopted for feature selection tasks [15, 16]. The proposed method in this paper shows an ensemble approach 'EnSNR' for feature (gene) selection. A genetic algorithm is used to build the classification model in the experiments. The filter approach has been described on many occasions: for example, Correlation-based Feature Selection [17] and Fast Correlation-based Feature Selection [18].

Recently, many researchers have employed the ensemble feature selection approach, instead of the more traditional approaches to feature selection, to improve algorithm classification accuracy [19-23]. Generally, ensemble-based feature selection approaches have two steps. Step 1: selecting subsets of informative features using different evaluation functions. Step 2: combining the features obtained from Step 1. In the proposed approach (ref: Step 1), two different feature evaluation functions are used to provide an input to the GA classifier: (1) Entropy and (2) SNR. Details of Entropy and SNR functions, are as follows:

3.1 'Entropy' Function

The 'Entropy' function is used to obtain a subset of features, from the 50,739 features in the Microarray dataset, for inputting to the genetic algorithm (see Section 4). The Entropy function is run 50,739 times to find the 'Feature Numbers' of those features with Entropy Score = 0 ('zero').

Before Entropy function calculations are performed, all of the '32 x 50,739' gene expression levels in the Microarray dataset (these are floating point numbers which can have a decimal component: the range of expression levels for a gene is between a minimum negative (- 11) level and a maximum positive (+16) level), are rounded to convert them to whole numbers. If a gene expression level does not have a decimal component, i.e. it is already a whole number, no conversion takes place. Also, if a dataset gene expression level is (0) 'zero' no conversion is necessary.

Rounding is performed, as follows: A positive floating-point number is rounded up to the next higher 'whole' number e.g. +10.01 is rounded up to +11; and so on. However, a negative floating-point number is rounded down to the next lower 'whole' number e.g. - 0.345 is rounded down to -1, and -1.15 is rounded down to -2; etc.

After rounding, gene expression levels are input to a Gene Expression Level (Rounded) look-up table, GEL(R). This table is only used for Entropy function calculations. The table consists of a matrix of 32 columns (for patient numbers) by 50,739 rows (for gene numbers). Each cell in the table contains a patient gene expression level (rounded) i.e. one of '28' (rounded) gene expression levels, including (0) 'zero'.

In the GEL(R) look-up table, gene expression levels (rounded) are referred to as integers. Each patient 'integer' expression level (for each feature (gene)) is assigned to a separate integer group. There can be up to 28 integer groups for each gene. Up to 32 patient numbers can be present in an integer group. The Entropy Score for a feature is calculated, as follows:

$$E(x_i) = \sum_b \left[\frac{n_b}{n_t} \left(-\sum_c \left(\frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b} \right) \right) \right] \quad (1)$$

Where:

- $E(x_i)$: Entropy score for feature (gene) number 'i' (1 to 50,739)
- 'b' : integer group number in an integer group set (i.e. cyclic progression of assigned integer groups within the range from 1 to 28) for feature (gene) number 'i'
- 'c' : Class-1 ('cancer') or Class-2 ('no-cancer')
- 'n_b' : number of patients in group 'b'
- 'n_t' : total number of patients in the dataset (32 patients)
- 'n_{bc}' : number of Class 'c' patients in integer group 'b'

There are 31 features which have an Entropy score of 'zero'.

3.2 Signal to Noise Ratio ('SNR') Function

The Signal to Noise Ratio 'SNR' function is used to obtain a subset of features, from the 50,739 features in the Microarray dataset. Features numbers (gene numbers) 1 to 50,739 (i.e. all feature/gene numbers) are input to the 'SNR' function. A separate SNR calculation is carried out for each feature, to find its SNR score; and the features are ranked. The SNR Score for a feature is calculated, as follows:

$$SNR(x_i) = \left| \frac{\overline{X_{i1}} - \overline{X_{i2}}}{SD_{i1} + SD_{i2}} \right| \quad (2)$$

Where:

$SNR(x_i)$: SNR score for feature (gene) number 'i' (i.e. 1 to 50,739)

$\overline{X_{i1}}$: average value of feature i (Class-1) i.e. average patient gene expression level for Class-1 ('cancer') patients, for gene i

$\overline{X_{i2}}$: average value of feature i (Class-2) i.e. average patient gene expression level for Class-2 ('no-cancer') patients, for gene i

SD_{i1} : standard deviation value of feature i (Class-1)

SD_{i2} : standard deviation value of feature i (Class-2)

The 31 features with the highest SNR scores are selected for the SNR subset.

Note: The number of SNR features selected (i.e. 31) is defined by the number of features with Entropy score 'zero'.

3.3 Ensemble ('EnSNR') Process

The Ensemble 'EnSNR' process, ref: the combination process mentioned in Step 2 (see details given earlier in this Section), is used to find the subset of features which are present in both Entropy and SNR subsets.

The contents of Entropy and SNR feature subsets are compared. Those features which are present in both subsets (14 features are found to be common to both subsets) become the 'EnSNR' subset. The EnSNR subset is used for creating an EnSNR dataset (which is used instead of the Microarray dataset) for the genetic algorithm input in experiment-4 (see Section 6).

3.4 Prior Related Works

Three prior related works associated with feature selection, are described below.

Saeys, Y., et al., [19] employ the Symmetrical Uncertainty (SU) and RELIEF algorithms for the univariate and multivariate approaches, respectively. The main purpose of Saeys's work is to investigate the use of ensemble feature selection techniques, where univariate and multivariate approaches are combined. Saeys's work deals with both the feature ranking and the correlation between features in an ensemble feature subset.

However, the EnSNR approach has some advantages over Saeys's approaches. For example, Entropy and SNR functions are used for obtaining the EnSNR feature subset using a ranking based feature selection method. These two evaluation functions are simple and only consume a small amount of computational time.

Xu, J., et al., [23] propose a NMICFS-PSO method where correlation-based feature selection (CFS), neighbourhood mutual information (NMI) and particle swarm optimization (PSO) are used for ensemble feature selection. Xu's work studies an ensemble method constructed from both filter and wrapper approaches. Then, gene ranking is performed. NMICFS-PSO consumes much more computational time than the proposed EnSNR method in this paper. Also, selected features from a wrapper approach would lead to a high degree of overfitting. In addition, EnSNR has a filter-based evaluation function. Therefore, features in the EnSNR subset are independent of the classification algorithm.

Ghosh, M., et al., [24] propose a 2-stage ensemble feature selection approach using microarray datasets. In their approach an ensemble feature subset is obtained using three ranking-based methods (ReliefF, chi-square, and symmetrical uncertainty). The top-n features of each method are selected: so, there are separate sets of ranked features for the ReliefF, chi-square, and symmetrical uncertainty functions. The union set of top-n features from these three methods is generated, together with three sets of top-n intersection features. A genetic algorithm (GA) is employed to find the optimal features from the union set and the intersection set, separately. The limitations imposed by the user-defined parameter (n) are a drawback of this approach. The EnSNR approach proposed in this paper deals automatically with the selection of the number of informative features to be processed, according to which microarray dataset is used.

4 GA - Based Classification Algorithm

The Genetic Algorithm (GA) is widely used for feature selection classification problems. A genetic algorithm, which employs the stochastic search method inspired by natural selection, provides improved prediction accuracy compared to other approaches. A genetic algorithm is ideal for solving problems with a large number of solutions, as in the experiments described in this paper, where there is a considerable quantity of data to process [25-27]. In the GA, classification and validation are employed to ensure the accurate analysis of gene expression level data coming from the Microarray dataset, with the lowest possible risk of miscalculation. The GA-Based Classification Algorithm proposed by Hengpraprohm [1] is used for the research described in this paper.

There are several steps involved in using a GA. A population of individuals (candidate solutions) is generated. Then, each individual is evaluated using the fitness function. GA operations (reproduction, crossover and mutation) are performed on individuals in the population pool to create new 'child' individuals. All new 'child' individuals are evaluated using the fitness function. Individuals having a high level of fitness survive and move on to the next generation for processing. The process is repeated until a stopping criterion is satisfied. The philosophy of the genetic algorithm is that 'child' individuals tend to inherit excellent characteristics/features from their parents; and survive to the next generation. The population evolves and provides the 'best' solution. In the experiments, the 'best' solution is the classification 'model'.

4.1 Dataset Information

The following information is available in the Microarray dataset. Relevant details, as appropriate, regarding how dataset information is used by the algorithm, are as follows:

Patient Identity: This is used to obtain Patient Number.

Patient Cancer Status: Breast Cancer Subtype (i.e. Luminal A, Luminal B or Triple negative), or normal (i.e. no-cancer). This information in the dataset is the patient cancer status, i.e. the 'known' condition mentioned previously, at the time when patient tissue samples were taken; and subsequently deposited on the microarray. In this work, patients are grouped together into two classes for classification purposes. A patient suffering from any of the three cancer subtypes given in the dataset is assigned Class 1 (cancer); and a patient with no sign of cancer is assigned Class 2 (no-cancer).

Gene ID: 11 alpha/ numeric characters. Gene ID is used in **Table 3**; and to obtain Gene Number.

Gene Symbol: Gene symbols consist of alpha numeric strings. They are only used in **Tables 3** and **4**. A gene symbol is a gene reference; or 'unknown' is given. 'Unknown' denotes an unknown function of a gene.

Gene Expression Level: The range of expression levels for a gene is between a minimum negative (-11) level and a maximum positive (+16) level. A negative value corresponds to an under expressed gene and a positive value corresponds to an over expressed gene: zero (0) is the normal 'mean' expression level.

In the experiments in this paper, patients are identified by a 'patient number'; and features (genes) by a 'gene number'.

Patient Number: 1 to 32. Patient numbers (derived from Patient Identities) are obtained from the Patient Number look-up table.

Gene Number: 1 to 50,739. Gene numbers (derived from Gene ID's) are obtained from the Gene Number look-up table.

4.2 Genetic algorithm parameters

Genetic algorithm parameters used in the experiments are defined in [1]. Parameter details are as follows:

Population Size (P): The population size is 100 chromosomes.

Chromosomes: Each chromosome consists of 20 chromosome elements. Elements are numbered 1 to 20 (element-1 is on the left-hand side of the chromosome). They are divided into two (2) groups. Elements 1 to 10 are assigned to Group-1. Elements 11 to 20 are assigned to Group-2. A feature (gene) number is present in each chromosome element.

Genetic algorithm operators

Reproduction Rate (R): 0.05 (i.e. 5%). This operator is used to pass on the five (5) strongest chromosomes to the 'next' generation (see details given in 'Genetic Algorithm processing' in this Section).

Crossover Rate(C): 0.90 (i.e. 90%). One (1) Point Crossover, employing the 'Tournament Selection' technique (tournament size = 5), is used in the experiments (see details given in 'Genetic Algorithm processing' in this Section).

Mutation Rate (M): 0.05 (i.e. 5%). One (1) Point Mutation, also employing the 'Tournament Selection' technique (tournament size = 5), is used in the experiments (see details given in 'Genetic Algorithm processing' in this Section).

Maximum Number of Generations (G): 1000.

Termination Conditions: There are two (2) termination conditions. When either of the termination conditions occurs, the algorithm classification process stops. Termination conditions are: 1). A prediction accuracy score of 100% is obtained; or 2). 1000 generations have been processed by the algorithm.

4.3 Genetic Algorithm processing

Processing starts with the generation of an 'initial' population.

Initial Population:

The algorithm puts 100 different randomly selected 'sets' of features (genes) into Chromosomes 1 to 100 to generate an 'initial' population; i.e. a total quantity of 2,000 features (genes). There are 20 feature (gene) numbers in a 'set'. All features (genes) have an 'equal opportunity' to be selected each time the Genetic Algorithm randomly chooses twenty (20) of them to populate these Chromosomes: feature numbers can appear more than once in each 'set'.

The features (genes) input to the algorithm in experiments 1 to 4 are as follows:

- *Experiment-1*: Features 1 to 50,739 from the Microarray dataset
- *Experiment-2*: Entropy 31-feature subset (these features have an Entropy score of (0) 'zero')
- *Experiment-3*: SNR 31-feature subset (these are the features with the highest SNR scores: the number of features in experiments 2 and 3 are the same)
- *Experiment-4*: EnSNR 14-feature subset (these are the 14 features which are present in both Entropy and SNR subsets)

Prediction accuracy calculations:

The prediction accuracy for the contents of Chromosome-1 is calculated by the algorithm. Gene expression levels for each of the twenty (20) gene numbers in the elements of Chromosome-1 are obtained from the microarray dataset. An expression level 'sum' is calculated (separately) for the genes in the two chromosome element groups (i.e. Group-1 and Group-2); and the two-expression level 'sums' are compared. If Group-1 sum > Group-2 sum, the algorithm prediction is Class 1 (cancer). However, if Group-1 sum ≤ Group-2 sum, the prediction is Class 2 (no-cancer).

Then, the chromosome prediction is compared with the 'known' cancer status (obtained from the dataset) for each patient (patients 1 to 32) and a prediction score calculated for Chromosome-1. If the chromosome's prediction (Class 1, or Class 2) is the same as the known cancer status (cancer, or no-cancer) for all of the 32 patients, the score is 100%; and the algorithm stops. However, if none of the predictions are correct, the score is 'zero'. Otherwise, if the score is somewhere in between, it is calculated as follows: e.g. if the algorithm prediction is correct for 20 patients, but incorrect for the remaining 12 patients; the score assigned is (20 ÷ 32 × 100) 62.5%.

The algorithm prediction is now calculated for Chromosomes 2 to 100 (in a similar manner to Chromosome-1). Chromosome prediction results (i.e. percentages) for Chromosomes 1 to 100 are input to a Chromosome Results look-up table. Then, chromosome results are 'ranked' and placed in a Chromosome Results (Ranked) look-up table.

However, algorithm processing terminates (stops) immediately if a prediction score of 100% (this is the ideal prediction score that the algorithm is seeking) is obtained during prediction score calculations.

Processing of the 'initial' population is now complete.

Population evolution

The 'initial' population is modified by three 'operators' (Reproduction, Crossover and Mutation) to obtain the first (1st) generation. Modifications are based on ranking levels in the Chromosome Results (Ranked) look-up table. These modifications are incorporated into the elements of Chromosomes 1 to 100 to form the 'new' population to be processed by the algorithm.

Reproduction:

The top five (5) chromosomes (ranking levels 1 to 5) given in the Chromosome Results (Ranked) look-up table for the 'initial' population, are copied into Chromosomes 1 to 5 (respectively) to provide the first (1st) generation gene number element contents for these chromosomes.

On completion of algorithm processing for the first (1st) generation, chromosome contents (for Chromosomes 1 to 5) are copied again, as described above, for the second (2nd) generation; and so on.

Crossover:

The Tournament Selection technique (tournament size = 5) is used to randomly select two (2) chromosomes from the Chromosome Results (Ranked) look-up table. By using one (1) Point Crossover, the crossover point is randomly selected. The crossover point is one of the twenty (20) elements in each of the selected chromosomes; and is the same element position in both of them. In each chromosome there are twenty elements. Element one (1) is on the left of the string.

Once the crossover point has been determined, the gene numbers in all elements to the right of the crossover point are swapped over (i.e. exchanged) between the two selected chromosomes. The crossover process is repeated another forty-four (44) times, as described above for the first pair of chromosomes selected. Crossover provides the gene number contents for a total of 90 'new' chromosomes for the 'next' generation to be processed by the algorithm.

Mutation:

The mutation operator randomly selects one (1) chromosome from the Chromosome Results (Ranked) look-up table, to be a parent, using the Tournament Selection technique (tournament size = 5). Then, the mutation point in the chromosome is randomly selected by the algorithm (the mutation point is one of the twenty (20) elements in the 'selected' chromosome). Random selection is then used by the algorithm to obtain a replacement gene number, which is input to the 'selected' chromosome element. The contents of all other elements (i.e. the other 19) in the selected chromosome remain unchanged.

The mutation process is repeated another four (4) times (as described above) to obtain the gene number contents for a total of five (5) 'new' (child) chromosomes for the 'next' generation to be processed.

Once Reproduction, Crossover and Mutation operations have been completed, the 'new' population of 100 chromosomes is ready for processing.

Processing of the 'new' population:

Algorithm processing continues with the 'Prediction accuracy calculations' described previously. Then, chromosome prediction results (i.e. percentages) for Chromosomes 1 to 100 are input (as before) into a Chromosome Results look-up table. Also, chromosome results are 'ranked' and placed in a Chromosome Results (Ranked) look-up table.

When processing of the 'first' generation is complete; Reproduction, Crossover and Mutation operators are used, as before, to provide the contents of 100 'new' chromosomes for the 'second' generation, as already described.

Algorithm processing continues in this manner until one of the termination conditions occurs i.e. 1000 generations have been processed, or a prediction accuracy score of 100% is obtained. On termination of algorithm processing, the single 'best' solution obtained is the classification 'model'. The 'model' is the 'top' ranked chromosome in the Chromosome Results (Ranked) look-up table.

The efficiency of the classification process is then checked using the 10-Fold Cross-Validation re-sampling procedure (see Section 5).

Fitness Function

The fitness function in the experiments is the training set prediction accuracy. Prediction accuracy is calculated as follows:

$$Prediction\ Accuracy = \frac{TP+TN}{N} \quad (3)$$

Where:

TP : number of true positive instances i.e. the number of correct 'cancer' case predictions by the algorithm

TN : number of true negative instances i.e. the number of correct 'no-cancer' case predictions by the algorithm

N : number of all instances i.e. the number of patients (28, or 29) in the training set

5 '10'-Fold Cross-Validation

The 10-Fold Cross-Validation re-sampling procedure [28] is used to test the efficiency of the classification process once algorithm processing has terminated (termination conditions are given in Section 4). This procedure is carried out in order to estimate how the classification 'model' (obtained on completion of algorithm processing) is expected to perform in the 'real world', when making predictions on data other than that coming from the microarray dataset used to train it.

To test the classification 'model', a 'training set' and 'test set' are employed. The 10-fold cross-validation re-sampling procedure is run 10 times.

Each run, which consists of ten (10) folds, is processed in the following manner. The set of observations (patients/ instances) for re-sampling is divided into 10 groups (or folds) of approximately equal size. In the experiments, there are 32 patients (instances) to be accommodated in the ten (10) groups. Group-1 (i.e. fold group-1) and group-2 (i.e. fold group-2) have four (4) members each. Groups 3 to 10 (i.e. fold groups 3 to 10) have three (3) members each. The training set has nine (9) fold groups: one (1) fold group is the test set. For the first cycle (i.e. fold) of the re- sampling procedure, fold groups 2 to 9 are the training set and fold group-1 is the test set. For the next cycle, fold group-1 and fold groups 3 to 10 are the training set (fold group-2 is the test set); and so on for the remaining eight cycles (folds) of the procedure.

At the beginning of the 10-fold cross-validation re-sampling procedure, the 32 patient numbers are shuffled to assign members (i.e. patient numbers) to fold groups 1 to 10. The assignment of patient number members to the ten (10) groups remains the same during the ten (10) folds. The training set is input to the genetic algorithm to generate a classification 'model'. There are 28, or 29, patient numbers in the training set, depending on the assignment of groups to the fold to be processed (see details given above).

The (ten) 10 folds (i.e. cycles 1 to 10) of the re-sampling procedure, are processed in the following manner.

- Fold-1(i.e. cycle-1): Fold groups 2 to 10 are the training set: group-1 is the test set. When algorithm processing of the training set terminates (termination conditions are given in Section 4), the 'best' result', i.e. the algorithm classification 'model' for the training set, is the 'top' ranked chromosome in the Chromosome Results (Ranked) look-up table (see 'Genetic algorithm processing' in Section 4).
- The training set classification model is then validated using the test set to obtain the prediction accuracy result for Fold-1. Algorithm processing time is logged in the system. Then, validation continues with the processing of the 'next' fold cycle.
- Fold-2: Fold group-1 and groups 3 to 10 are the training set: group-2 is the test set. The 'best' result (i.e. the algorithm classification 'model') for 'fold-2' is processed in a similar manner to 'fold-1'. Validation continues; with the fold group number of the test set incrementing by 1, each time, at the beginning of each 'new' fold (i.e. cycle) of the process.
- Fold-10: Fold groups 1 to 9 are the training set: group-10 is the test set. When the algorithm terminates, the 'best' result (i.e. the algorithm classification 'model') for fold-10 is processed in a similar manner to 'fold-1'.

For each run of the 10-fold cross-validation re-sampling procedure, the average prediction accuracy for the ten (10) folds is given in **Table 1** and **Figure 2**. Algorithm processing time for each run is given in **Table 2**. The average algorithm processing time for the ten runs is given in **Figure 3**.

6 Experimental Set - Up

The GA-Based Genetic Algorithm proposed by Hengpraprohm [1] is used for classification purposes in all four experiments.

Features input to the algorithm in each experiment are as follows:

- Experiment-1: All 50,739 features (genes)
- Experiment-2: The Entropy subset of 31 features
- Experiment-3: The SNR subset of features (also, 31)
- Experiment-4: The EnSNR subset of 14 features

Experiments 1 to 4 are carried out using a modern laptop computer (CPU 2.60 GHz, RAM 8 GB).

The data source for the experiments is a Microarray Dataset: this is downloaded to the laptop computer from the web site (see Section 2). The dataset format is normalized CSV data.

The four experiments are entirely separate. There is no sharing of data between them.

Two look-up tables, for use in the experiments, are generated from the microarray dataset, as follows:

- Patient Number (1to 32) look-up table
- Gene Number (1 to 50,739) look-up table

The results of experiments 1 to 4, i.e. average prediction accuracy and algorithm processing time, are given in Section 7 for each 'run' of the 10-fold cross-validation re-sampling procedure.

Experiment-1 (All Features):

Features (gene numbers) 1 to 50,739 are input to the genetic algorithm in this experiment. Prediction accuracy is calculated by the algorithm for the 'initial' population; and for each generation. The algorithm stops running after 1000 generations have been processed; or a prediction accuracy of 100% (this is the ideal condition which the algorithm is seeking) is achieved during the processing of training set data.

Experiment-2 ('Entropy' Features):

A Gene Expression Level (Rounded) look-up table is prepared. This table is only used for Entropy function calculations. Features 1 to 50,739 are input to the 'Entropy' function. A separate Entropy calculation is carried out for each feature to find its Entropy score. Features having Entropy score = 0 ('zero') are placed in an Entropy look-up table (i.e. 31 features with Entropy score 'zero'); and an Entropy dataset is created. The 31 features in the Entropy dataset are input to the genetic algorithm. Prediction accuracy is calculated by the algorithm for the 'initial' population; and for each generation.

The algorithm stops running after 1000 generations have been processed, or a prediction accuracy of 100% is achieved during algorithm processing of training set data.

Experiment-3 (Signal to Noise Ratio 'SNR' Features):

Features 1 to 50,739 are input to the 'SNR' function. A separate SNR calculation is carried out for each feature, to find its SNR score. The features are ranked. The (31) features with the highest SNR scores (note: the number of SNR features is equal to the number of Entropy features with Entropy score 'zero') are placed in an SNR look-up table; and an SNR dataset is created. The 31 features in the SNR dataset are input to the genetic algorithm. Prediction accuracy is calculated by the algorithm for the 'initial' population; and for each generation.

The algorithm stops running after 1000 generations have been processed, or a prediction accuracy of 100% is achieved during algorithm processing of training set data.

Experiment-4 (Ensemble 'EnSNR' Features):

Features 1 to 50,739 are input to the 'Entropy' function. A separate Entropy calculation is carried out for each feature to find its Entropy score. Features having Entropy score = 0 ('zero') are placed in an Entropy look-up table (31 features with Entropy score 'zero' are found when the Entropy function is processed).

Features 1 to 50,739 are input to the 'SNR' function. A separate SNR calculation is carried out for each feature to find its SNR score; and the features are ranked. The 31 features with the highest SNR scores (note: the number of SNR features is equal to the number of Entropy features with Entropy score 'zero') are placed in an SNR look-up table.

The contents of the Entropy and SNR look-up tables are compared. Those features which are present in both tables (14 features are found to be common to both tables) become the 'EnSNR' subset of features; and an EnSNR dataset is created. The 14 Features in the EnSNR dataset are input to the genetic algorithm. Prediction accuracy is calculated by the algorithm for the 'initial' population; and for each generation. The algorithm stops running after

1000 generations have been processed, or a prediction accuracy of 100% is achieved during algorithm processing of training set data.

7 Experimental Results and Discussion

Results of experiments 1 to 4 are given in this Section. For each experiment, **Table 1** shows the prediction accuracy obtained for each run of the 10-Fold Cross-Validation re-sampling procedure. The average prediction accuracy for the 10 runs, together with the standard deviation, is given in the bottom row of **Table 1**. A graphical representation of the average prediction accuracy obtained in each experiment is shown in **Figure 2**.

Table 1: Prediction accuracy percent (%).

10-Fold Cross-Val. Run No.	**Genetic Algorithm Input Data**			
	Experiment-1: All 50,739 Features	**Experiment-2: Entropy 31 Features**	**Experiment-3: SNR 31 Features**	**Experiment-4: EnSNR 14 Features**
1	50.83	**82.5**	78.33	78.33
2	51.67	74.17	65.83	**91.67**
3	69.17	81.67	71.67	**85.83**
4	50	67.5	65.83	**84.17**
5	45.83	75.83	66.67	**88.33**
6	49.17	66.67	70.83	**97.5**
7	69.17	**85**	69.17	80
8	45	72.5	62.5	**87.5**
9	50.83	**89.17**	67.5	88.33
10	54.17	75	64.17	**87.5**
Avg. ± SD.	53.58 ± 8.64	77.00 ± 7.41	68.25 ± 4.54	**86.92 ± 5.47**

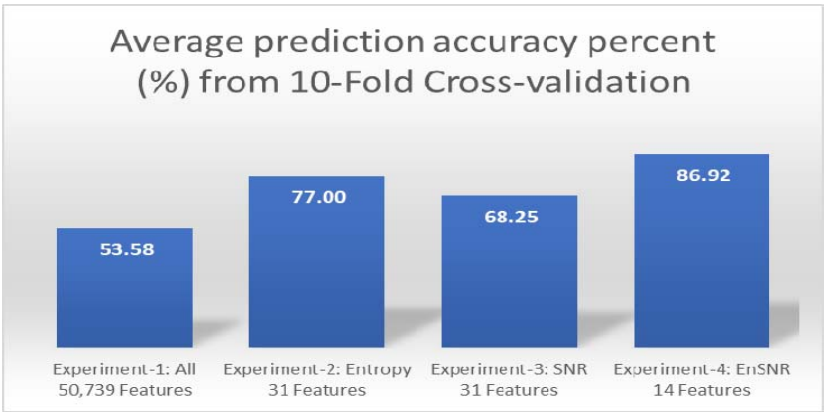

Figure 2. Average prediction accuracy percent (%) from 10-Fold Cross-validation

The average prediction accuracy score obtained for the EnSNR feature subset, in Experiment 4, is 86.92%. This result is better, in 7 out of 10 runs of the cross-validation procedure (i.e. Run Nos. 2, 3, 4, 5, 6, 8 and 10 in **Table 1**), than that obtained when the experiments are carried out with the full set of features (50,739); the Entropy subset of 31 features; or the SNR subset of 31 features.

In addition, the average prediction accuracy of the proposed EnSNR approach (Experiment-4) is statistically compared, using 'pair t-tests' with a level of 0.05, with the average prediction accuracy obtained for 'All Features' (Experiment-1), 'Entropy' (Experiment-2) and 'SNR' (Experiment-3). The results of the three 'pair t-tests' show that the average prediction accuracy of the EnSNR approach is significantly higher than that obtained from experiments 1, 2 and 3.

For each experiment, **Table 2** shows the processing time for each run of the cross-validation re-sampling procedure. Average processing time for each run, together with standard deviation, is given in the bottom row of

Table 2. **Figure 3** provides a graphical representation of the average processing time obtained in experiments 1 to 4.

The EnSNR subset of features shown in **Table 3** could be used in data processing to make more accurate breast cancer predictions, than using the full set of 50,739 features; or the 'Entropy', or 'SNR', subsets of features. When the EnSNR subset is used, in addition to increased prediction accuracy, there is also a reduction in algorithm processing time, for the learning process to generate the classification 'model' than when using the full set of features (see **Figures 3**). Processing time is similar when the EnSNR, Entropy, or SNR subsets are input to the algorithm (see **Table 2**). However, when the 'EnSNR' subset is used (instead of the 'Entropy', or 'SNR', subsets), it can be seen from **Figure 3** that there is a small saving in average processing time for the ten (10) 'runs' of the 10-fold cross-validation re-sampling procedure.

Table 2: Genetic Algorithm processing time (secs.).

10-Fold Cross-Val. Run No.	**Genetic Algorithm Input Data**			
	Experiment-1: All 50,739 Features	**Experiment-2: Entropy 31 Features**	**Experiment-3: SNR 31 Features**	**Experiment 4: EnSNR 14 Features**
1	28.33	4.75	4.71	**4.64**
2	26.68	4.23	4.30	**3.52**
3	28.74	4.27	4.84	**3.32**
4	26.85	4.24	5.13	**3.73**
5	28.62	4.29	4.33	**3.65**
6	28.48	4.33	4.30	**3.82**
7	28.97	4.30	4.43	**3.11**
8	28.59	4.21	4.28	**3.55**
9	29.07	4.18	4.27	**3.70**
10	27.60	4.30	4.24	**3.72**
Avg. ± SD.	28.19 ± 0.85	4.31 ± 0.16	4.48 ± 0.30	**3.68 ± 0.40**

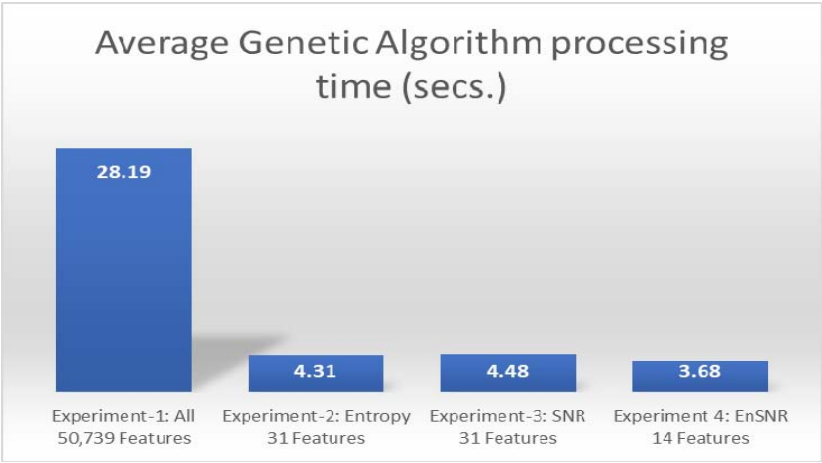

Figure 3. Average Genetic Algorithm processing time (secs.)

In addition, the classification accuracy result of the proposed EnSNR approach compares favourably with the latest work in breast cancer classification proposed by López-Cabrera et al. [29]. López-Cabrera et al. employ Convolutional Neural Networks (CNN) to classify three classes of breast cancer (i.e. two cancer subtypes and no-cancer) from digital mammography data. There are two CNN architectures: a 3-class architecture and an architecture with two CNNs in series. Prediction accuracies are 86.05% (for the 3-class architecture) and 88.20% (for the architecture with two CNNs in series); while the proposed EnSNR approach in this paper gives 86.92%.

This paper is concerned with the selection of features to be used for breast cancer diagnosis. However, there are four features (genes) in the EnSNR 14 feature subset that are also associated with other types of cancer, or a

cancer related process. Details of these four genes, and the relevant research papers in which they are described, are given in **Table 4**.

Table 3: 'EnSNR' feature subset (14 features).

Feature No.	Microarray Dataset Information		Entropy Score	SNR Score
	Gene ID (Microarray)	Gene Symbol (Microarray)		
1	A21P0010122	Unknown	0	3.759635
2	A33P3254460	Unknown	0	3.135592
3	A23P340171	**TP53AIP1**	0	2.795258
4	A24P602871	Unknown	0	2.719541
5	A33P3313125	Unknown	0	2.662858
6	A33P3497352	**GRIA4**	0	2.416312
7	A21P0011734	LOC100505851	0	2.402573
8	A23P126248	Unknown	0	2.386455
9	A33P3266739	FLJ41200	0	2.228636
10	A21P0006274	Unknown	0	2.222591
11	A33P3214988	Unknown	0	2.210122
12	A33P3289121	Unknown	0	2.207569
13	A33P3370094	**MME**	0	2.190308
14	A33P3254606	**WDR62**	0	2.163969

Table 4: Cancer Related Genes (ref: **Table 3**).

Gene Symbol (Microarry)	Cancer Type/ Cancer Process: and Research Paper Reference
TP53AIP1	Prostate cancer related [26]
GRIA4	Cancer cell proliferation related [27]
MME	Lung cancer related [28]
WDR62	Ovarian cancer related [29]

8 Conclusions

This paper proposes an ensemble feature selection approach, 'EnSNR', for gene selection for breast cancer microarray data classification. This method follows an ensemble-based feature selection trend. 'Entropy' and 'SNR' evaluation functions are employed to find the relevant informative features. After that the selected feature subset is given to the GA classifier. The dataset in these experiments contains 50,739 features (genes) for each of 32 patients. The well-known 10-fold cross-validation procedure is employed to measure the prediction accuracy of the classification model. In the EnSNR approach, the Entropy subset of features (i.e. 31 features with an Entropy score of 'zero') are combined with the 31 highest-ranked SNR features (note: the number of SNR features is equal to the number of Entropy features with Entropy score 'zero'), to obtain the Ensemble, EnSNR, subset of 14 features. All features present in the EnSNR subset appear in both Entropy and SNR subsets. The experiments show that the Ensemble, EnSNR, approach selects a smaller number of features (i.e. 14) than the Entropy and SNR approaches. Also, better performance in terms of prediction accuracy is obtained.

Because the EnSNR features appear in both Entropy and SNR subsets (they are the 'best' features obtained using two different selection methods) this gives more confidence in the breast cancer prediction accuracy (rather than using the higher number of features selected by Entropy, or SNR). The Entropy and SNR functions provide some additional advantages for the EnSNR approach. For example, the number of features in the EnSNR subset is not user-defined (i.e. the EnSNR subset is generated automatically depending on which microarray dataset is used); and the operation of the EnSNR function is independent of the type of classification algorithm employed. Also, only a small amount of processing time is required to generate the EnSNR feature subset.

Further action: since only 32 patient/gene profiles were examined in the experiments (and all of them came from female patients), it is necessary to carry out further research with gene profiles coming from both male and female patient tissues (as breast cancer can also occur in men) before drawing any firm conclusions about the suitability of the four genes (referred to in Section 7) in the wider spectrum of cancer diagnosis (other than them

only being used in the subset of features for breast cancer prediction). It would also be interesting to carry out experiments with EnSNR using gene databases to try to improve the prediction accuracy still further; and to possibly discover additional genes related to cancer.

Acknowledgements

The authors are indebted to Nakhon Pathom Rajabhat University (NPRU) for support of the research described in this paper.

References

[1] Hengpraprohm, S., (2013, June). "GA-Based Classifier with SNR Weighted Features for Cancer Microarray Data Classification," International Journal of Signal Processing Systems, Vol.1, No.1, pp. 29-33.

[2] Supoj Hengpraprohm and Prabhas Chongstitvatana, "Feature Selection by Weighted-SNR for Cancer Microarray Data Classification", International Journal of Innovative Computing Information and Control (IJICIC), Vol. 5, No. 12(A), pp. 4627-4635, December 2009.

[3] Dziuda, D. M., (2010). Data Mining for Genomics and Proteomics: analysis of gene and protein expression data. New Jersy: Wiley & Sons.

[4] Babu, M. M., (2004). Introduction to microarray data analysis. In: G. RP, ed. Computational Genomics: Theory and Application. Norwich,: Horizon Press, pp. 225-249.

[5] Junjie Fu, William Allen, Amy Xia, Zhuofan Ma and Xin Qi, (2014). “Identification of biomarkers in breast cancer by gene expression profiling using human tissues”, Genomics Data, Vol.2, pp. 299-301.

[6] George, G. V. S., Raj, V. C., (2011). Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification Using Gene Expression Profile. International Journal of Computer Science & Engineering Survey. Vol.2(3), pp. 16-26.

[7] Saeys, Y., Inza, I. and Larranaga, P, (2007). A review of Feature Selection Technique in Bioinformatics. Bioinformatics, Vol.23(19), pp. 2507-2517.

[8] Lui, H., Motoda, H., (1998). Feature Selection for Knowledge Discovery and Data Mining. Massachusetts: Kluwer Academic.

[9] Liu, H., and Motoda, H., (2007). Computational methods of feature selection. CRC Press.

[10] Liu, H., Motoda, H., Setiono, R., and Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. FSDM 10, Vol.4(13).

[11] José Menezes, Giordano Cabral, Bruno Gomes and Paulo Pereira. (2019). Feature Learning with Multi-objective Evolutionary Computation in the generation of Acoustic Features. Inteligencia Artificial Vol.22(64), pp.14-35.

[12] Jungjit, S. (2016). New Multi-Label Correlation-Based Feature Selection Methods for Multi-Label Classification and Application in Bioinformatics (Doctoral dissertation, University of Kent,)

[13] Dash, M., and Liu, H., (2003). Consistency-based search in feature selection. Artificial intelligence 151, Vol.155(176).

[14] Kudo, M., & Sklansky, J., (2000). Comparison of algorithms that select features for pattern classifiers. Pattern recognition, Vol.33(1), pp.25-41.

[15] Hong, J. H., & Cho, S. B. (2006). Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognition Letters, Vol.27(2), pp.143-150.

[16] Dessi, N., and Pes, B., (2009). An evolutionary method for combining different feature selection criteria in microarray data classification. Journal of Articial Evolution and Applications 2009, p.3.

[17] Hall, M. A., (1999) Correlation-based feature selection for machine learning.

[18] Yu, L., and Liu, H., (2003), Feature selection for high-dimensional data: A fast correlation-based filter solution. In ICML, Vol. 3, pp. 856-863.

[19] Saeys, Y., Abeel, T., & Van de Peer, Y. (2008, September). Robust feature selection using ensemble feature selection techniques. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 313-325.

[20] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, Vol.26(3), pp.392-398.

[21] Opitz, D. W. (1999). Feature selection for ensembles. AAAI/IAAI, pp.379 - 384.

[22] Tsymbal, A., Puuronen, S., & Patterson, D. W. (2003). Ensemble feature selection with the simple Bayesian classification. Information fusion, Vol.4(2), pp.87-100.

[23] Xu, J., Sun, L., Gao, Y. and Xu, T. (2014). An ensemble feature selection technique for cancer recognition. Bio-medical materials and engineering Vol.24(1), pp.1001-1008.

[24] Ghosh, M., Adhikary, S., Ghosh, K.K., Sardar A., Begum, S. and Sarkar, R. (2019). Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & biological engineering & computing. Vol.57(1). pp.159-176.

[25] Eiben, A. E., & Smith, J. E. (2003). Introduction to evolutionary computing, Vol. 53.

[26] Freitas, A. A. (2002). Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer Science & Business Media.

[27] Freitas, A. A. (2005). Evolutionary algorithms for data mining. In Data mining and knowledge discovery handbook, Springer, Boston, MA, pp. 435-467.

[28] Witten, I. H., Frank, E. and Hall, M.A. (2011). Data Mining Practical Machine Learning Tools and Techniques. 3rd Ed. Burlington: Morgan Kaufmann.

[29] José Daniel López-Cabrera, Luis Alberto López Rodríguez and Marlén Pérez-Díaz, (2020). Classification of Breast Cancer from Digital Mammography Using Deep Learning. Inteligencia Artificial Vol.23(65), pp.56-66.

[30] Manuel Luedeke , Irina Coinac , Carmen M. Linnert, Natalia Bogdanova, Antje E. Rinckleb, Mark Schrader, Walther Vogel, Josef Hoegel, Andreas Meyer, Thilo Dörk, Christiane Maier, (2012. March) "Prostate Cancer Risk Is not Altered by TP53AIP1 Germline Mutations in a German Case-Control Series", PLoS ONE Vol.7(3): e34128, https://doi.org/10.1371/journal.pone.0034128

[31] Hella Luksch, Ortrud Uckermann, Andrzej Stepulak, Sandy Hendruschk, Jenny Marzahn, Susanne Bastian, Christian Staufner, Achim Temme and Chrysanthy Ikonmidou, (2011, October) "Silencing of Selected Glutamate Receptor Subunits Modulates Cancer Growth", Anticancer Research, vol. 31(10), pp. 3181-3192.

[32] Katharina Leithner, Christoph Wohlkoenig, Elvira Stacher, Jörg Lindenmann, Nicole A Hofmann, Birgit Gallé, Christian Guelly, Franz Quehenberger, Philipp Stiegler, Freyja-Maria Smolle-Jüttner, Sjaak Philipsen, Helmut H Popper, Andelko Hrzenjak, Andrea Olschewski and Horst Olschewski, (2014, January) "Hypoxia increases membrane metallo-endopeptidase expression in a novel lung cancer ex vivo model – role of tumor stroma cells", BMC Cancer, pp.14:40.

[33] Yu Zhang, Yan Tian, Jing-Jing Yu, Jie He, Jia Luo, Sai Zhang, Cen-E Tang and Yi-ming TaoEmail, (2013). "Overexpression of WDR62 is associated with centrosome amplification in human ovarian cancer", Journal of Ovarian Research, Vol.6:(55), https://doi.org/10.1186/1757-2215-6-55