Cancer Subtype Identification Pipeline: A Classifusion Approach

Utkarsh Agrawal
School of Computer Science
University of Nottingham
Nottingham, UK
psxua@nottingham.ac.uk

Daniele Soria
School of Computer Science
University of Nottingham
Nottingham, UK
daniele.soria@nottingham.ac.uk

Christian Wagner
School of Computer Science
University of Nottingham
Nottingham, UK
christian.wagner@nottingham.ac.uk

***Abstract*— Classification of cancer patients into treatment groups is essential for appropriate diagnosis to increase survival. Previously, a series of papers, largely published in the breast cancer domain have leveraged Computational Intelligence (CI) developments and tools, resulting in ground breaking advances such as the classification of cancer into newly identified classes - leading to improved treatment options. However, the current literature on the use of CI to achieve this is fragmented, making further advances challenging. This paper captures developments in this area so far, with the goal to establish a clear, step-by-step pipeline for cancer subtype identification. Based on establishing the pipeline, the paper identifies key potential advances in CI at the individual steps, thus establishing a roadmap for future research. As such, it is the aim of the paper to engage the CI community to address the research challenges and leverage the strong potential of CI in this important area. Finally, we present a small set of recent findings on the Nottingham Tenovus Primary Breast Carcinoma Series enabling the classification of a higher number of patients into one of the identified breast cancer groups, and introduce Classifusion: a combination of results of multiple classifiers.**

Keywords—Breast Cancer; Classification methods; Consensus Classification (Classifusion); Consensus Clustering

I. INTRODUCTION

Breast Cancer is one of the most common cancer types in women [1]-[3]. It creates the particular difficulty of determining the suitable treatment option, among a number of complex treatment choices. Recent studies [1]-[5] have solved this problem by identifying treatment classes using Computational Intelligence (CI), and dividing patients among them. This development evolved in the past two decades is scattered among several works, as detailed below.

Eisen et al. [6] introduced hierarchical clustering combined with the visual study of dendrograms to analyse the relation among the gene expression. The work inspired the use of clustering techniques in the cancer profile with respect to genes data [7]-[9]. Perou et al. [9] used the hierarchical clustering algorithm on gene expression data identifying four molecular classes: *luminal, HER2, basal* and *normal*. A subsequent study further subdivided the luminal into three, namely *A, B* and *C* [10]. Sotiriou et al. [11] carried forward the research partitioning the basal group into two, and eliminating the normal group, thus obtaining a total of six groups. With advancements in technology, an alternative approach developed is to use immunocytochemistry on formalin fixed paraffin embedded patient tumour sample. Researchers have applied multiple protein biomarkers with relevance to breast cancer, investigating the clinical importance of breast cancer classes [1]-[5],[12],[13].

To address the stability issue of the classification from clustering techniques, Monti et al. [12] introduced consensus clustering, which considers the consensus across multiple clustering algorithms. The idea behind consensus clustering is to capture the robustness of sampling variability of the clusters formed. If this measure increases, the degree of confidence of the overall classification structure also increases. Kellam et al. [13] refers to consensus clustering as Clusterfusion, which takes into account the results of different clustering algorithms and creates robust clusters depending on the consensus across the results of multiple clustering algorithms.

Contrasting and combining the results of multiple clustering algorithms is significant to test the stability of the clusters formed and of the proposed classification. Similarly, to test the stability of the classification outcome we introduce 'Classifusion', which takes the results of different classification algorithms and generates a set of robust classes based upon the consensus of the results of each algorithm. This allows us to infer level of confidence in regard to certain samples belonging to a particular class.

The proposed pipeline to analyse the breast cancer patient's data is introduced in this paper and explained in a step-by-step manner with reference to prior work, in particular [1]-[5]. Analysis of immunohistochemical data and different clustering techniques has led to seven clinical groups for breast cancer data composed of protein biomarkers which were collected using tumour samples [1],[3]. The work was followed by a fuzzy rule based classification algorithm (FuzzyQSBA) combined with a novel class assignment method [2]. The method classified patients in one of the seven classes, but in the process few patients remained unclassified. The aim of this work is to develop a cancer subtype identification pipeline and present the most recent developments i.e. apply classifiers to the previously 'Not Classified' breast cancer patients to assess similarities with the well-defined classes and thus to assign these patients to a specific class. As the ground truth data does not exist for this type of cancer-based datasets, the resulting classification results are validated by employing a series of clustering techniques. Two specific datasets are used as part of this paper, the first being the original breast cancer dataset [1] and the second called 'fuzzy membership values dataset'

produced as a result of the FuzzyQSBA algorithm mentioned above [2].

The structure of the paper is as follows: Section II presents the Pipeline, which contains the step-by-step (pipeline) architecture for the cancer subtype identification. Section III explains the methodology used. Section IV contains the results and the steps of Classifusion. In Section V the discussion of the results is presented. Section VI concludes the paper with future directions of research.

II. Pipeline

A. Division of patients into core classes

To assist prognosis in the context of breast cancer treatment, Soria et al. 2010 analysed a breast cancer panel composed of 25 protein biomarkers, using several clustering algorithms and a consensus approach [1]. The work inspected five clustering techniques namely Hierarchical, K-means, Partitioning around medoids, Adaptive Resonance Theory and Fuzzy C-means. The clustering was followed by the computation of validity indices to validate the number of clusters and to tackle the issue of cluster stability. Consensus among the clusters was evaluated through statistical and visualisation methods such as Principal Component Visualisation and Boxplots, and by two algorithms namely Artificial Neural Network and Rule Extraction. By establishing the consensus the work identified six novel cancer subtypes, and classified the patients among these core classes. The process identified prime biological classes*: luminal, HER2* and *basal* (*luminal* and *basal* containing distinct subclasses), shown in the breast cancer classification structure in Fig.1. The procedure classified 61.6% of patients into one of the core classes while the remaining 38.4% were assigned to a mixed class. Further analysis was required to minimise the misclassification.

B. Reduction of the biomarkers to an essential set

In a subsequent study, Green et al. 2013 [3] studied boxplots of the complete dataset and individual classes (described in Soria et al. [1]) and reduced the biomarkers on which the classes were derived from 25 to 14. The panel was reduced by omitting the biomarkers which had identical overall distributions across multiple classes. The study focused on the reduction of the number of biomarkers which were sufficient to obtain good classification. Using a Naive Bayesian supervised classification approach as used in Soria et al. [4], the set of biomarkers was further reduced to contain the ten 'most important' ones. The exhaustive search using the naive Bayes classification to choose the best combination of ten biomarkers out of 14 was done to also reduce the costs of clinical tests.

The work also studied the class characterisation by analysing the immunohistochemical profiles of the patients. Class 6 (HER2) showed heterogeneity in the expression of the hormone receptor which led to division of HER2 into two subclasses HER2+/ER+ and HER2+/ER- shown in Fig. 2. The stable clusters validate the core classes (by Soria et al. [1]) and forms basis of further study (by Soria et al. [2]).

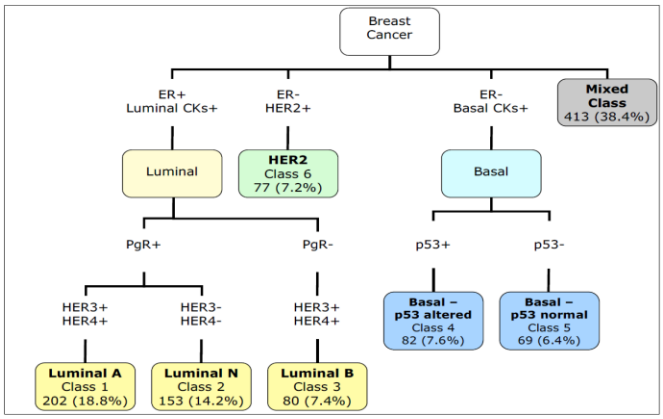

Fig. 1. Breast Cancer Classification structure by Soria et al. [1].

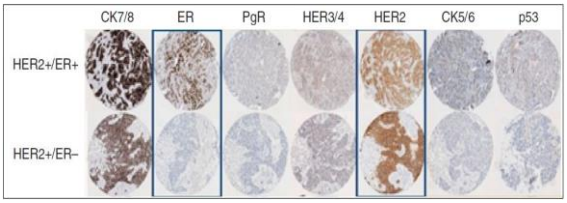

Fig: 2. Subdivision of HER2+ in two subclasses HER2+/ER+ and HER2+/ER- based on immunohistochemical profiles by Green et al. [3]

C. Division of new patients into core classes by learning the pattern

The analysis was further extended in the work by Soria et al. [2], which explored the use of fuzzy methodologies to generate a rule set which can classify a large number of patients into one of the identified classes. The work makes use of a data-driven fuzzy rule based system (FuzzyQSBA) for classifying the breast cancer patients. The FuzzyQSBA [2] itself is similar to the algorithm developed in Rasmani et al. 2009 [5]. This latter study makes use of 663 classified patients reported in Soria et al. [1].

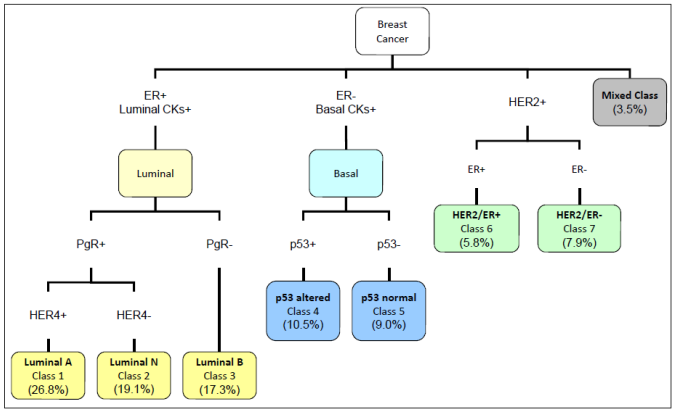

Fig: 3. Breast Cancer Classification structure by Soria et al. [2].

The algorithm developed in [2] was applied to a well characterised dataset of 1,073 breast cancer patients. The patients were classified into one of the identified seven classes using fuzzy iteration method, similar to the algorithm developed in the work Green et al. [3]. The FuzzyQSBA algorithm learns the class-labelled breast cancer dataset by generating one classification rule per class. The algorithm

assigns a class membership value to each patient for each of the seven classes. A novel class allotment algorithm is developed which assigns patients into one of the seven classes. In doing so, only 38 out of 1,073 patients were misclassified. The class distribution is represented in Fig. 3. The classification rules were validated on a new breast cancer dataset of 238 patients. Only 12 patients were assigned to 'Not Classified', giving an excellent classification outcome. The complete research pipeline described so far is represented in the flowchart diagram reported in Fig. 4.

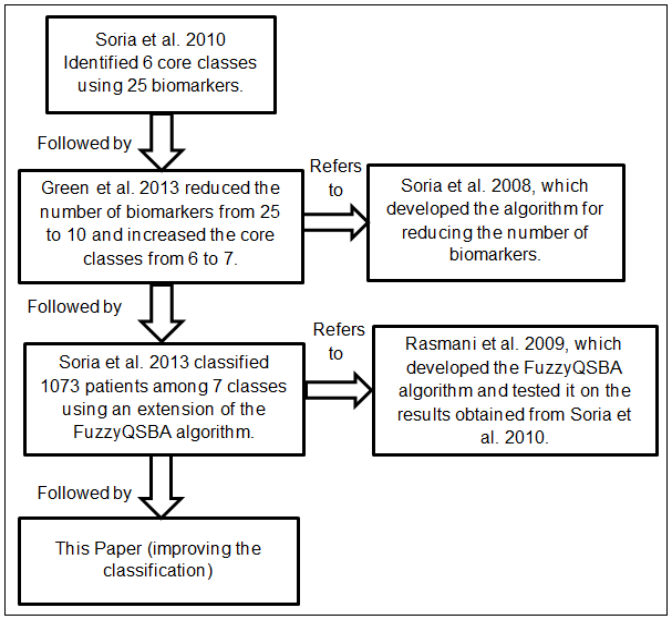

Fig. 4. Flowchart of the research.

1) FuzzyQSBA algorithm:

The FuzzyQSBA algorithm is a fuzzy rule induction method based on fuzzy subsethood values i.e. they represent the degree to which a fuzzy set is a subset of another fuzzy set. The aim of using this technique is to replace crisp weights in the Weighted Subsethood-Based Algorithm (WSBA) [2] by fuzzy quantifiers. The main difference of Quantifier Subsethood-Based Algorithm (QSBA) [2] compared to WSBA lies in the interpretation of the inference between weights/quantifiers with the linguistic terms. In QSBA both the quantifiers and the linguistic terms are fuzzy sets. It offers us the flexibility to use t-norm operators to interpret the rule.

The FuzzyQSBA algorithm consists mainly of these four steps: 1) Divide the dataset into subgroups. 2) Calculate fuzzy subsethood value. 3) Calculate relative weights. 4) Create fuzzy rules. Examples of fuzzy rules are as follows:

WSBA - "IF A is (A1 OR 0.09A2) AND B is (B2 OR 0.2B3) AND C is (C1) THEN Output is D".

FuzzyQSBA - "IF A is ((almost all)A1 OR (a little)A2) and B is ((almost all)B2 OR (almost a quarter of)B3) AND C is ((almost all)C1) THEN Output is D".

The goal of the above steps was to provide the likelihood of membership of each patient in each class. After obtaining these values, the class assignment algorithm was applied. For class assignment, the two highest membership values were chosen. If the difference between the top two values was greater than a certain threshold, then the patient was assigned to the class with maximum membership. If the difference was below the threshold and they both belong to same class family then the class with maximum values was chosen. Otherwise the patient was not classified.

D. Classification

1) Artificial Neural Network (ANN):

The Artificial Neural Network (ANN) [4],[14],[15] aims to classify the data by using the backpropagation learning method. The input nodes receive a finite number of inputs from the dataset, and pass the data using the synapse to the hidden layer. Now using another set of synapse the data is passed to the output layer, where a weighted sum is calculated and compared with a threshold. The network learns by back propagating the error and readjusting the weights. The learning process continues until the error reaches the error threshold. One of the major reasons behind ANN performing the tasks effectively is the nature of network, which enables multiple neurons collaterally in the direction of solving the problem.

2) Nearest Neighbour:

The Nearest Neighbour algorithm [16] begins by considering that data are clustered in their respective class. It follows a greedy approach by calculating the distance of the test data with each classified data. It then assigns to the class to the test data on the basis of minimum distance i.e. the test data is assigned to the class with the shortest distance. The algorithm is represented in Fig. 5 where x_j is one of the 'Not Classified' breast cancer patients. ω_1, ω_2 and ω_3 represent the identified classes in which the patient is intended to be classified.

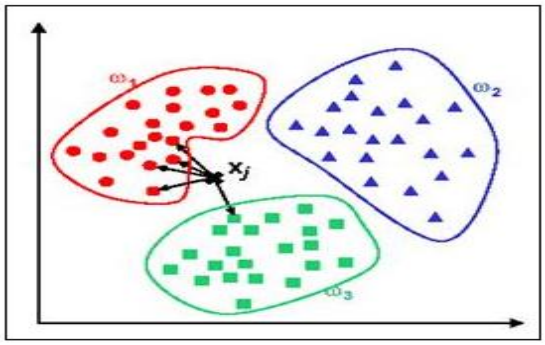

Fig. 5. Nearest Neighbour representation [20].

3) Classifusion:

We introduce a new approach termed as 'Classifusion' i.e. a consensus across all the classifiers. This is done to provide robustness to the result. An object is assigned to a class when it is classified in the same class by all the classification algorithms. In this work the two classifiers classify the patients into one of the identified seven classes.

E. Clustering

1) Hierarchical clustering:

Hierarchical clustering [1],[4],[6],[9] is a technique used for cluster analysis. There are two ways to form clusters, the first being agglomerative which follows a bottom up approach;

and the second being divisive, which follows a top down approach. In this work the agglomerative strategy is used to form clusters.

The algorithm's structure is as follows: 1) Assign each item to a cluster, i.e. if there are 'n' observations then we have 'n' clusters. 2) Find the closest cluster pair based on the pre-specified distance. 3) Merge the two clusters and calculate the average of them, which will be the representation of the new cluster formed. 4) Compute the distance between the new cluster and the old clusters. 5) Repeat the steps 2-4 until all the observations come under one single cluster.

2) K-Means:

K-means clustering [1],[17] works by splitting 'n' observations into 'k' clusters. The algorithm works by grouping the observations with similar properties in one group, in such a way that it differs from the clusters having different properties. The algorithm's steps are: 1) Start by choosing 'k' random cluster centres. 2) Calculate the distance of all the 'n' observations with each centre. 3) Assign the observations to the centre with the minimum distance. 4) Recalculate the centres by taking the average of each newly formed cluster. 5) Repeat steps 2-4 until no new assignment is made.

3) K-Medoids:

K-Medoids clustering, also known as Partitioning Around Medoids (PAM) [1],[19] aims to partition 'n' observations into 'k' clusters. It is related to both the K-means and the medoid shift algorithms, and works by minimising the sum of pairwise dissimilarities. The algorithm's steps are: 1) Randomly select 'k' cluster centres (medoids) among 'n' observations. 2) Assign each point to the closest medoid. 3) For each medoid point 'm' and non-medoid point 'o', swap 'm' and 'o' and re-compute the cost. 4) If the new cost is more than the previous cost, undo the swap. 5) Repeat steps 2-4 until all the swaps are made.

F. Datasets:

In this paper, two specific datasets are considered:

1) Breast Cancer Dataset:

The Breast Cancer Dataset consists of 1,073 patients from the Nottingham Tenovus Primary Breast Carcinoma Series with operable stages I, II and III. Immunohistochemical reactivity [18] for 25 proteins was determined using tissue microarray (tumour samples). Among the available information, we consider the same ten important protein markers suggested by [3] and run our test on them. The same reduced set of ten protein biomarkers has been used in the recent study by Soria et al. [2] to classify patients in seven clinical classes using a data driven FuzzyQSBA algorithm.

2) Fuzzy membership value dataset:

Soria et al. [2] used the FuzzyQSBA method together with a class assignment algorithm to classify the breast cancer patients in one of the seven breast cancer classes. The algorithm was run on the reduced set of ten protein biomarkers from the breast cancer dataset. The FuzzyQSBA algorithm [2],[5] developed uses continuous fuzzy quantifiers to create the ruleset, which assigns a class membership value to each patient for each of the seven classes. The advantage of using a fuzzy rule based model is its ability to represent the information in the form of if-then rules. This mechanism of representation is human understandable and combines several rules to generate global behaviour of the system.

Fuzzy membership values in Fig. 6 are the membership values obtained after implementing FuzzyQSBA on the ten protein biomarkers, numbered from 1 to 7, and form what we call the 'fuzzy membership value dataset'. The column Class represents one of the seven identified classes in which the patient is classified. ER, PgR … MUC1 represents the ten protein biomarkers from the Breast Cancer Dataset on which the FuzzyQSBA algorithm was applied.

III. Methodology

In this work two algorithms namely: Artificial Neural Network (ANN) [4],[14],[15] and Nearest Neighbour [16], were used for classification of 'Not Classified' patients. The underlying strategy is based on the following: 1) to classify 'Not Classified' patients in one of the identified classes, 2) to increase accuracy by considering multiple classification methodologies. To establish a strong and accurate classification the algorithms were run on the breast cancer dataset consisting of protein markers and on the fuzzy membership values of each patient for the seven classes. The classification results were validated by applying three clustering techniques namely K-means, Hierarchical Clustering and K-Medoids. The major steps followed were:

1) First the ANN classification algorithm was run on the 1,073-patients breast cancer dataset, where 1,035 were classified in one of the seven classes. The ten protein biomarkers data of 1,035 patients were used to train the 3-layer neural network. The network was tested with multiple numbers of hidden nodes, optimal number being 20. After the network was trained by updating its weight, the updated network was tested on the 38 'Not Classified' patients. Similar steps were adopted with the Fuzzy membership value dataset. The labelled data of seven class membership values for the 1,035 patients was used to train the neural network. After the network learned and updated its weight the updated network was tested on the 38 'Not Classified' patients. The classification outcomes of both the datasets are explained in the 'Results after Classification' segment of the Results section.

					Fuzzy Membership values							
Patient	ER	PgR	...	MUC1	1	2	3	4	5	6	7	Class
1625	280	0	...	235	0.20	0.15	0.90	0.10	0.10	0.55	0.20	3
2879	0	0	...	130	0.10	0.15	0.20	0.65	0.20	0.10	0.10	4
3932	200	295	...	200	0.85	0.10	0.20	0.15	0.20	0.10	0.05	1

Fig. 6. Tuples of data set produced after applying the FuzzyQSBA algorithm.

TABLE I. RESULTS AFTER CLASSIFICATION

Patient No.	Ann original (1)	Ann mem (2)	Nearest neigh original (3)	Nearest neigh mem (4)	1 & 3	2 & 4	Among all 4
2	1	1	1	1	1	1	1
3	4	4	4	4	4	4	4
7	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5
14	4	3	4	4	4	0	0
24	5	5	5	5	5	5	5
25	4	4	4	4	4	4	4
32	5	5	5	5	5	5	5
35	4	4	4	4	4	4	4
38	4	4	4	4	4	4	4

2) The Nearest Neighbour algorithm calculates the distance of 38 'Not Classified' with all the labelled 1,035 patients. It assigns the class to 38 patients based on minimum distance. The same strategy was used in both the breast cancer dataset and the fuzzy membership values dataset. In the breast cancer dataset the distance was calculated among ten protein biomarkers, and in the fuzzy membership values dataset the distance was calculated among the membership values. The classification outcome for both the datasets is explained in the 'Results after Classification' segment of the Results section.

3) The classification results were verified by changing the parameters of the classifiers, followed by the consensus between the classification algorithms (Classifusion).

4) K-means [1],[17] was run on both the breast cancer dataset and the Fuzzy Membership value dataset. Results of the K-Means clustering on the Fuzzy membership value dataset are explained in the 'Results after Clustering' segment of the Results section but the results of the K-means applied to the ten protein biomarkers of the breast cancer dataset are not reported as they were not informative and therefore dropped from further analysis.

5) Hierarchical [1],[4],[6],[9] and K-medoids [1] clustering algorithms were run on both the breast cancer dataset and the Fuzzy Membership value dataset. Outcome of both the algorithms on both the datasets are explained in the 'Results after Clustering' segment of the Results section.

6) Consensus between the results of the different clustering algorithms was studied, and compared with the classification outcomes.

IV. RESULTS

A. *Result after Classification:*

Classifusion works by generating a list which checks the output class of an instance for each algorithm. If the class is the same throughout, then the instance is assigned to the consensus class, otherwise the instance is considered to be misclassified. The classifusion results are shown in the columns '1 & 3', '2 & 4' and 'Among all 4' of Table I. Abbreviations of the table headers are as follows:

'Ann original': ANN algorithm applied to the 10 protein biomarkers data with 20 hidden layer nodes and 20,000 iterations of weight updating. 'Ann mem': ANN algorithm applied on the fuzzy membership values dataset. 'Nearest neigh original': nearest neighbour algorithm applied to the 10 protein biomarkers data. 'Nearest neigh mem': nearest neighbour algorithm applied on the fuzzy membership values dataset. '1 & 3': consensus among the ANN and nearest neighbour algorithm applied to the 10 protein biomarkers values (classifusion). '2 & 4': consensus (classifusion) among the ANN and nearest neighbour algorithm applied to the fuzzy membership values. 'Among all 4': consensus (classifusion) among the ANN and nearest neighbour algorithm for both fuzzy membership values and 10 protein biomarkers values data sets.

Table I shows the classification results for ten randomly selected patients out of 38. Out of 38 patients, 31 had consensus results in case of '2 & 4' i.e. by combining the classification results classification of Fuzzy membership values dataset. In case of consensus of classification of breast cancer dataset, i.e. '1 & 3', the number reduced to 24. The consensus of classification for both the datasets ('Among all 4') further reduced down to 18 out of 38. It can be noted that the fuzzy membership values dataset generated better results. One

TABLE II. RESULTS AFTER CLUSTERING

Patient No.	K-medoids orig (1)	K-medoids mem (2)	Hierarchical clustering orig (3)	Hierarchical clustering mem (4)	K-means mem (5)	Among all 5 (1&2&3& 4&5)
2	1	1	1	1	1	1
7	5	5	5	5	5	5
8	4	4	4	4	4	4
14	2	4	2	4	4	0
16	1	1	1	1	1	1
21	4	5	4	5	5	0
22	3	3	3	3	3	3
26	4	5	4	4	5	0
35	4	4	4	4	4	4
38	4	4	4	4	4	4

explanation for the difference in numbers might be that the fuzzy membership values dataset is pre-processed, i.e. the dataset is the result of FuzzyQSBA algorithm.

A number of observations can be drawn after the classification process: firstly, when the algorithms were run on the membership values dataset patients were either classified into classes corresponding to the highest membership value or to the second highest membership value, but no patient was classified to class 6 or class 7. The membership values for class 6 and class 7 were a constant value for all the 38 patients. In addition, when the algorithms were run on the ten protein biomarkers similar results to those of the membership values were obtained, i.e. no patient was classified to class 6 or class 7. The reason for this was that the values of HER2 in the original 1,073 data were either 0 or 300. If the HER2 values were 300 then the patient belonged to either class 6 or class 7. If the value was 0 then the patient belonged to one of the remaining five classes or was misclassified. For all the 38 patients under investigation in this work, the values of HER2 were 0. So the patient would never belong to class 6 or class 7. We decided to omit class 6 and class 7 in further clustering analyses of 'Not Classified' breast cancer patients for the reasons explained above.

B. Result after Clustering:

Individual results (for ten randomly selected patients) after each clustering algorithm along with consensus across all the methods are reported in Table II. Abbreviation of the table headers are as follows:

'K-medoids orig': K-Medoids clustering algorithm applied on the ten original protein biomarkers. 'K-medoids mem': K-Medoids clustering algorithm applied on the fuzzy membership values. 'Hierarchical clustering orig': Hierarchical clustering algorithm applied on the ten protein biomarkers. 'Hierarchical clustering mem': Hierarchical clustering algorithm applied on the fuzzy membership values. 'K-means mem': K-Means clustering algorithm applied on the fuzzy membership values.

As reported in the classification discussion above we eliminated the possibility of cluster 6 and cluster 7, and reduced the number of clusters in which the patients could be grouped to five. Columns (1) to (5) of Table II represent the outcome of K-medoids, Hierarchical clustering and K-means clustering applied on the breast cancer dataset and on the fuzzy membership value datasets. Out of 38 patients, 20 had consensus results across all the five methods. The classification result classified 18 out of 38 patients in one of the seven classes. Out of 20 patients from clustering, 18 patients are exactly the same as those from the classification results. The consensus of clustering and classification results therefore classifies 18 patients into one of the identified classes.

K-medoids and Hierarchical clustering produced the same results with respect to each other by generating similar clusters. They classified the same 20 patients in one of the identified classes for both the datasets, and performed slightly better than the classification by assigning more patients into one of the classes. The results obtained by applying K-means on the breast cancer dataset were not informative and were dropped from further investigation. But, the results obtained on the fuzzy membership values dataset were similar to those produced by K-medoids and Hierarchical clustering i.e. they classified the same 20 patients into their respective classes.

K-means clustering algorithm performed poorly on the breast cancer dataset, but it produced good results in case of the fuzzy membership values dataset. This difference could be due to the fact that the fuzzy membership values are processed values and have been obtained after applying FuzzyQSBA on original breast cancer dataset.

V. Discussion

After applying clustering and classification techniques on the 38 'Not Classified' patients from the Nottingham breast cancer series [2], 18 of the patients can now be assigned to one of the seven classes. These patients show strong affinity with the original classes 4 and 5. Among 18 patients, six can be classified in class 4, 11 in class 5 and the one remaining in class 1. Fig. 7 shows the boxplots of the original 1,073 breast cancer patients where the y-axis represents the H-score and the x-axis represents the ten protein markers.

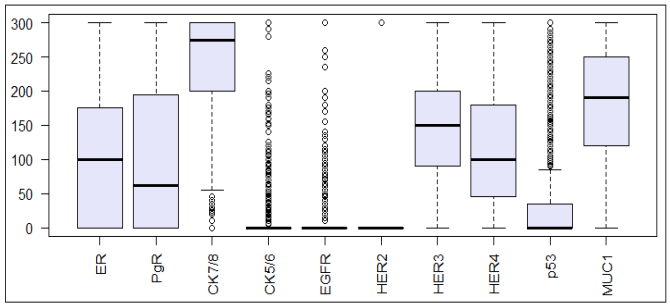

Fig. 7. Boxplot of 1073 patients.

Fig. 8 compares the boxplots of the original classes 4 and 5 with those from the current work. It can be seen that previous and new methodologies generate similar boxplots, validating the new results. The 'MUC1' biomarker has a different mean and a different range of values in the new classes 4 and 5. The mean is high in both the new classes. Similar behaviour is observed for the protein 'CK7/8' i.e. the mean is high and the range is different in both the new classes 4 and 5. The difference in the values of both the biomarkers could be the reason for misclassification of these patients in the original work by Soria et al. [2]. The strong affinity for class 4 and class 5 is validated by the range and the mean of the remaining eight biomarkers.

VI. Conclusion

In this work we have presented a step-by-step pipeline for cancer subtype identification with a case study from the breast cancer disease. The pipeline forms the basis for the appropriate diagnosis to increase survival of cancer patients. The work further improves the classification accuracy of the FuzzyQSBA result by classifying 18 more patients into one of the seven core breast cancer classes. This paper introduces classifusion i.e. a combination of results from multiple classifiers to make the classification more robust.

In the future we plan to present the results to clinical experts for their interpretation and we will continue working on the development of the pipeline and the 'Classifusion' approach.

Acknowledgment

The authors are thankful to the Nottingham Breast Cancer Pathology group at the Nottingham University Hospitals – NHS Trust for collecting and providing the data used in this study.

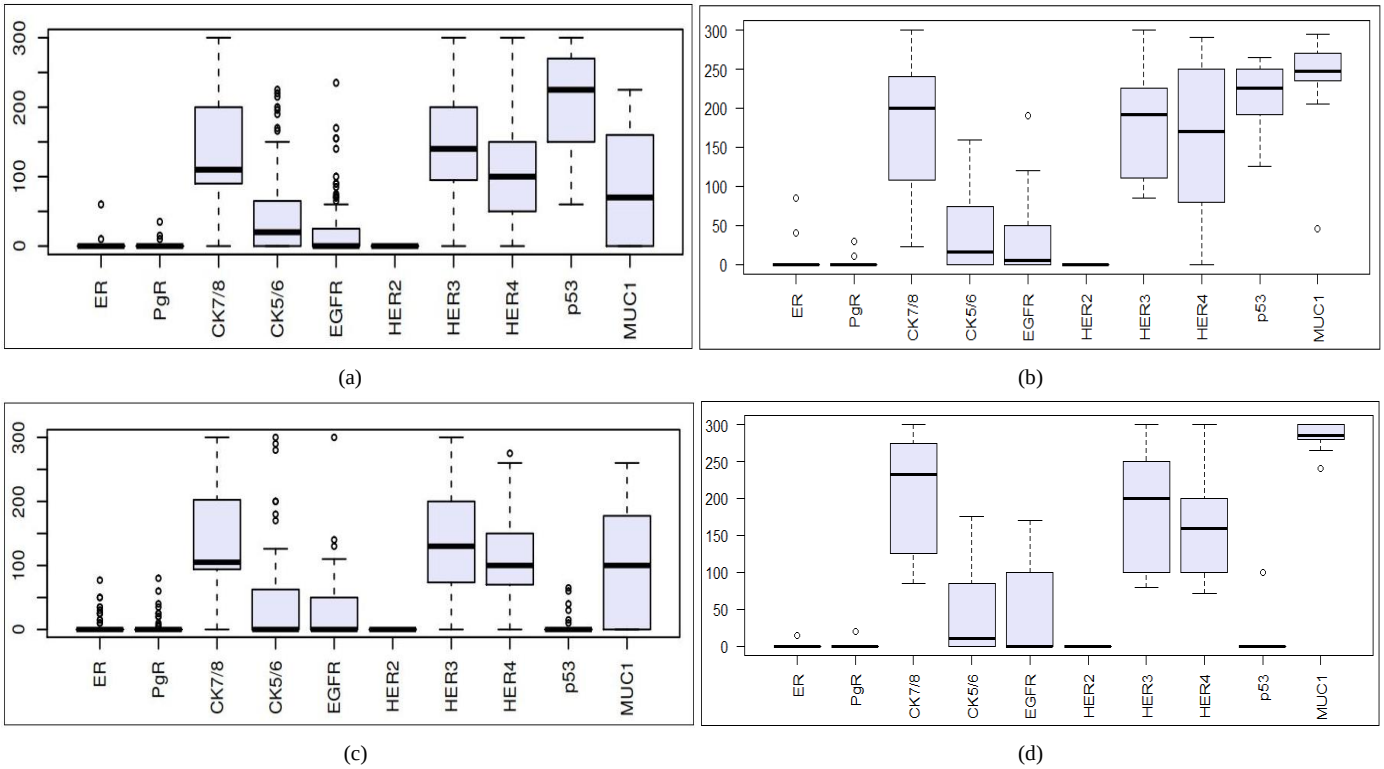

Fig. 8. Boxplots of patients (a) Class4 of original classification (b) Class4 of new methodology (c) Class5 of original classification (d) Class5 of new methodology

REFERENCES

[1] D. Soria et al., "A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients", Computers in biology and medicine 40.3 (2010): 318-330.

[2] D. Soria et al., "A quantifier-based fuzzy classification system for breast cancer patients", Artificial intelligence in medicine 58.3 (2013): 175-184.

[3] A. R. Green et al., "Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers", British journal of cancer 109.7 (2013): 1886-1894.

[4] D. Soria, J. M. Garibaldi, E. Biganzoli, and I. O. Ellis., "A comparison of three different methods for classification of breast cancer data", In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on, pp. 619-624. IEEE, 2008.

[5] K. Rasmani, J. M. Garibaldi, Q. Shen, and I. O. Ellis, "Linguistic rulesets extracted from a quantifier-based fuzzy classification system", In . FUZZ-IEEE 2009. IEEE International Conference on, pp. 1204-1209. IEEE, 2009.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[7] M. Bittner et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling", Nature, 406:536–540, 2000.

[8] L. J. V. Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer", Nature 415.6871 (2002): 530-536.

[9] C. M. Perou et al., "Molecular portraits of human breast tumours", Nature 406.6797 (2000): 747-752.

[10] T. Sorlie et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", Proc Natl Acad Sci USA 98.19 (2001): 10869-74.

[11] C. Sotiriou et al., "Breast cancer classification and prognosis based on gene expression profiles from a population-based study", Proc. Natl. Acad. Sci. USA 100 (2003) 10393–10398.

[12] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resamplingbased method for class discovery and visualization of gene expression microarray data", Machine. Learning. 52 (2003) 91–118.

[13] P. Kellam, L. Xiaohui, M. Nigel, O. Christine, S. Stephen Swift, and T. Allan, "Comparing, contrasting and combining clusters in viral gene expression data", In Proceedings of 6th workshop on intelligent data analysis in medicine and pharmocology, pp. 56-62. 2001.

[14] R. Polikar, "Pattern Recognition", Wiley Encyclopedia of Biomedical Engineering, 2006.

[15] J. Khan et al, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature medicine 7.6 (2001): 673-679.

[16] M. Sarkar, and T. Y. Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem", Proceedings of the AMIA Symposium. American Medical Informatics Association, 2000.

[17] U. Agrawal, S. K. Roy, U. S. Tiwary, and D. S. Prashanth., "K-means clustering for adaptive wavelet based image denoising", In Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in, pp. 134-137. IEEE, 2015.

[18] S. E. Elsheikh et al, "Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome", Cancer research 69.9 (2009): 3802-3809.

[19] H. S. Park, and C. H. Jun, "A simple and fast algorithm for K-medoids clustering", Expert Systems with Applications 36.2 (2009): 3336-3341.

[20] L. L. Hong, W. C. Heng, Y. T. Fui and K. H. Meian, "A Review of Nearest Neighbor-Support Vector Machines Hybrid Classification Models", Journal of Applied Sciences, 2010, Volume: 10, Issue: 17, pages: 1841-58.