Combining Clustering and Classification Ensembles: A Novel Pipeline to Identify Breast Cancer Profiles

Utkarsh Agrawal[a], Daniele Soria[b], Christian Wagner[a], Jonathan Garibaldi[a], Ian O. Ellis[c,d], John M.S. Bartlett[e,f], David Cameron[e], Emad A. Rakha[c,d], Andrew R. Green[c]

[a]*School of Computer Science, The University of Nottingham, Nottingham*
[b]*School of Computer Science and Engineering, University of Westminster, London*
[c]*Nottingham Breast Cancer Research Centre, Division of Cancer and Stem Cells, School of Medicine, Nottingham City Hospital, University of Nottingham, Nottingham*
[d]*Breast Institute, Nottingham University Hospitals NHS Trust, Nottingham*
[e]*The Institute of Genetics and Molecular Medicine, Edinburgh Cancer Research Centre, University of Edinburgh, Western General Hospital, Edinburgh*
[f]*Transformative Pathology, Ontario Institute for Cancer Research, MaRS Centre, Toronto*

Abstract

Breast Cancer is one of the most common causes of cancer death in women, representing a very complex disease with varied molecular alterations. To assist breast cancer prognosis, the classification of patients into biological groups is of great significance for treatment strategies. Recent studies have used an ensemble of multiple clustering algorithms to elucidate the most characteristic biological groups of breast cancer. However, the combination of various clustering methods resulted in a number of patients remaining unclustered. Therefore, a framework still needs to be developed which can assign as many unclustered (i.e. biologically diverse) patients to one of the identified groups in order to improve classification. Therefore, in this paper we develop a novel classification framework which introduces a new ensemble classification stage after the ensemble clustering stage to target the unclustered patients. Thus, a step-by-step pipeline is introduced which couples ensemble clustering with ensemble classification for the identification of core groups, data distribution in them and improvement in final classification results by targeting the unclustered data. The proposed pipeline is employed on a novel real world breast cancer dataset and subsequently its robustness and stability are examined by testing it on standard datasets. The results show that by using the presented framework, an improved classification is obtained. Finally, the results have been verified using statistical tests, visualisation techniques, cluster quality assessment and interpretation from clinical experts.

Keywords: Ensemble Clustering, Ensemble Classification, Class level fusion, Refining cluster results, Breast Cancer, Pipeline

Email addresses: psxua@nottingham.ac.uk (Utkarsh Agrawal), d.soria@westminster.ac.uk (Daniele Soria)

1. Introduction

Determining the number of groups in a dataset with no class labels is a very challenging task [1]. One such problem is to assist breast cancer prognosis by determining the most suitable treatment option among a number of complex treatment choices available [2], [3], [4]. This can be answered only by fully understanding the biological characteristics of the disease for each individual patient. Therefore, the identification of biological groups is extremely important to tailor and choose the ideal treatment. To tackle this breast cancer heterogeneity, computational intelligence methods have been developed to facilitate personalised breast cancer treatment and support clinical decision making by predicting the patient's treatment outcome. Thus, recent studies [5], [4], [6] have utilised clustering algorithms to classify patients into biological groups.

Clustering algorithms [7] are chosen according to the problem, for example, in breast cancer studies, generally the cluster algorithms considered are centroid and connectivity based [8], [9]. Previously, Eisen et al. [8] combined hierarchical clustering with the visual study of the dendrograms to determine the association between the gene expressions. Using clustering methods on breast cancer profiles, Perou et al. [10] recognised four distinct molecular classes of breast cancer based upon gene expression data: *HER2*, *basal*, *luminal* and *normal*. Extending the work, Sorlie et al. [11] subdivided the Luminal group into three: *A*, *B*, and *C*. Being uncertain of Luminal C, in a subsequent work [12] they divided basal into two and discarded the normal group and Luminal C, thus attaining an aggregate of five breast cancer groups.

A bigger problem lies in algorithm selection, as different algorithms produce different results. Thus, employing several algorithms is advantageous to provide a level of confidence to the final grouping. Kellam et al. [13] used this idea and introduced 'Clusterfusion', which takes into account the results of all the different clustering algorithms present in the ensemble. Monti et al. [14] used consensus clustering in combination with re-sampling techniques to identify biological groups. Similarly, Chen et al. [15] introduced an ensemble clustering method to predict survival rate of cancer patients when working with discrete features. The algorithm used Partitioning Around Methods recursively to construct a dissimilarity matrix to be used by Hierarchical Clustering Algorithm to obtain final distribution.

In the past decade, Soria et al. [16], [2], [3] used the clusterfusion approach on data from immunohistochemistry on formalin-fixed paraffin embedded patient tumour samples, an alternative approach to gene expression profiling, for better identification of breast cancer biological groups. The patients from Nottingham Tenovus Primary Breast Carcinoma Series were divided among six breast cancer classes, similar to previous works. In a subsequent work by Green et al. [17], the biomarkers panel was reduced from 25 down to a set of 'ten most important' biomarkers. Criteria for class membership were defined using the expression of a reduced set of 10 proteins able to identify key molecular classes. The reduced set was obtained using the association between these breast cancer classes with clinicopathological factors and patient outcome. This work also studied the class characterisation of breast cancer patients by analysing their biomarker profiles. It further suggested the division of class 6 (HER2 positive) into two subclasses.

Using the reduced biomarker panel with the need for refinement of class 6, Soria and colleagues [3] divided the set of 1,073 Nottingham breast cancer patients among seven biological classes. However, the methodology left behind a number of unclassified patients. Researchers have employed multiple relevant protein biomarker to large numbers of cases to elucidate breast cancer classes using ensemble clustering methods [2], [3], [17], [18], [10], but a framework still needs to be developed which can assign as many unclustered patients to one of the identified groups in order to improve final classification results. Thus, we introduce the idea of using an extra layer of classifiers after the ensemble clustering layer [19] to improve the final classification.

The first aim of this work is to classify the unclustered data obtained during ensemble clustering. For this purpose, we construct a novel general pipeline focusing on the identification of the optimal number of clusters, and then patients are distributed among the groups using ensemble clustering. Subsequently, a consensus voting ensemble classifier layer has been introduced to increase the number of data in one of the identified groups by targeting the unclustered patients remaining after the ensemble clustering. We verify the data grouped after ensemble classification by comparing it with the data clustered after ensemble clustering, using statistical (Mann-Whitney-Wilcoxon) [20] and visualisation (boxplots) tests. We also verify the cluster quality using the Davies-Bound index [21]. The results show that with the addition of ensemble classification, the pipeline improves the ensemble clustering results.

To classify unlabelled datasets (e.g. Breast Cancer Dataset) a number of clustering [8] and ensemble clustering methods [13, 14, 15] have been proposed. Recently, a combination of clustering and classification algorithms have attracted attention. Chakraborty et al. [22] proposed EC3: Combining Clustering and Classification, employing clustering to learn supplementary constraints (e.g., if two objects are clustered together, it is more likely that the same label is assigned to both of them) to support classification algorithms in order to classify data across unlabelled datasets. In another set of works by Zhang et al. [23], the authors used a combination of clustering and classification for unlabelled data in network traffic classification. The classification framework proposed in this paper is built on different settings to the ones aforementioned, and the idea is to use clustering algorithms to form core groups and subsequently classify more data by learning the pattern of these core groups.

The second aim of the work is to address the robustness of the pipeline by testing it on multiple datasets. Thus, this novel pipeline introduced in this work is operated on a real world dataset from the Edinburgh Breast Cancer Series [6] and validated on Standard Datasets from the Machine Learning UCI repository [24]. The results present an improved classification for all the datasets and show the generality of the pipeline for application in non-medical domains.

The structure of the paper is as follows: Section 2 provides background information on methods used in the work. Section 3 introduces the step-by-step methodology (including a process flowchart) for the pipeline. Section 4 explains the experimental settings and the results for the application on all the datasets. Section 5 presents the discussion of the results and finally section 6 concludes the paper with future research.

2. Background

This section presents a brief description of the clustering algorithms, classification algorithms and ensemble methods used in the pipeline.

2.1. K-means (with HCA)

The K-means clustering algorithm is one of the most commonly used clustering algorithms which works by assigning the data points among k clusters [25]. The steps followed in the algorithm are: 1) Randomly select k number of cluster centres. 2) Compute the Euclidean distance of each observation from all the cluster centres. 3) Assign each observation to the cluster with the minimum distance from the centre. 4) Re-evaluate the cluster centres by considering the means of the respective data points. 5) Repeat steps two to four until no new allocation occurs.

K-means suffers from the disadvantage of relying on the random selection of cluster centres for each initial run, leading to the formation of dissimilar clusters each time. To tackle this problem multiple solutions have been proposed, and in this work we employ agglomerative Hierarchical Clustering Algorithm (HCA) [26] on the dataset before the K-means algorithm. The tree formed from the HCA is pruned, and the points are passed as the initial cluster centres (now fixed centres) to the K-means [2]. The K-means with centres from HCA is the new Modified K-means (with HCA) algorithm used in this work.

2.2. Partitioning Around Medoids (PAM)

Partitioning Around Medoids (PAM) is a clustering algorithm which characterises clusters by their medoids (also called the centres) [27]. It works similarly to the K-means, but in contrast it selects real data points as medoids. The algorithm works in two phases: 1) The build phase, in which a set of 'k' medoids among the 'n' observation points is selected. Then each remaining data point is assigned to one of the closest medoids based on the Euclidean distance. 2) The swap phase, in which each non medoid point is replaced with one of the medoid points and the distance is recalculated. If the new distance is greater in comparison with the previous distance, the swap is reversed, otherwise the process is repeated until all the swaps are performed.

2.3. Ensemble Clustering

Ensemble clustering methods merge results of multiple clustering algorithms to form core groups and have been successfully used in the domain of breast cancer [2], [3], [28], [13]. Several researchers have concentrated on contrasting and combining results of different clustering algorithms using re-sampling techniques, combination of outputs, probabilistic methods, pairwise similarity, etc. [29]. Among these multiple ensemble clustering methods, a Consensus Clustering approach [30] i.e. Clusterfusion is chosen because it provides robustness, stability and improved solution by increasing the degree of confidence of the groups formed [31].

Validity Index	Rule
Marriot	$max_k((d[k+1]-d[k])-(d[k]-d[k-1]))$
Calinski	$min_k((d[k+1]-d[k])-(d[k]-d[k-1]))$
Scott	$max_k(d[k]-d[k-1])$
TraceW	$max_k((d[k+1]-d[k])-(d[k]-d[k-1]))$

Table 1: Rules associated with validity indices

2.4. Cluster Validity Indices

Cluster validity indices are metrics used for determining the goodness of clusters by assessing the quality of clustering results [1]. The validity indices are either based on compactness of data in a cluster or on the separation of the clusters from one another. A specific rule is associated with each of the indices, as shown in Table 1, which specifies the optimal number of clusters k by ranking them. Thus, by using these measures in a dataset the most stable cluster size can be determined. Several validity indices have been proposed in the literature, but in this work four commonly used validity indices d are used: Marriot, Calinski, Scott and TraceW [1], [2].

2.5. Artificial Neural Networks

The Artificial Neural Network (ANN) algorithm [32], [33] employed in this work is also called shallow Deep Neural Network, with backpropagation learning method. The input layer receives a set of inputs from the dataset and transfers the information onto the hidden layer using synapses. Thus, with the help of more synapses the information is passed onto the third layer or the output layer, where a weighted sum is computed and passed to an activation function. The value received after the activation function is compared to a threshold and the error is calculated. The error is back-propagated and all the weights are updated. The network learns in the process of weight update and continues until the error is below the 'error threshold'. The advantages of using ANN are its flexibility, the capability to model highly complex models and the non-parametric nature.

2.6. Nearest Neighbour

The Nearest Neighbour algorithm is a memory based non-parametric model which is employed to classify the data points to the nearest group [34]. It starts by computing the distance of a test sample with all the data which have been classified in their respective groups. The test sample is assigned to the nearest group based on the distance matrix calculated in the previous step. The advantages of the Nearest Neighbour over other classification algorithms are its efficient memory handling, as it keeps track of all the clustered data, and its successful use in the domain of breast cancer [34, 35, 36].

2.7. Ensemble Classification

Ensemble Classification is a broad term for the methods combining multiple classification algorithms to improve predictive performance of the system [20, 37]. Previously, several researchers [38, 39, 40, 41] have used classifier

ensemble to perform classification and discussed its superiority over single classifiers [42]. Bashir et al. [43] have used ensemble classifier for breast cancer diagnosis and suggest that ensemble methods mimic human reasoning, as they consider multiple opinions before making a final decision. Many combination techniques exist such as boosting, bagging, stacking, etc. [44]. Among a number of decision level fusion methods present in the literature, unanimous voting was chosen, as the real world problem in this work needs high confidence, being the division of Breast Cancer patients into biological classes. The unanimous decision on a test sample 'x' could be mathematically represented as:

$$class(x) = mode\{C_1(x), C_2(x), ...C_n(x)\}, \tag{1}$$

where n is the number of classifiers in the ensemble and $C_n(x)$ is the final decision of the n-th classifier in the ensemble.

3. Methodology for the Pipeline

While combining the clustering algorithm results using ensemble methods, a number of data are left unclustered. Thus, in this section, we introduce a new step-by-step framework for the identification of groups and efficient data distribution within them. Researchers have focused on making ensemble clustering combination as good as possible, so to leave the smallest number of data unclustered [2, 29]. Unlike other approaches, the presented pipeline deals directly with the unclustered data by introducing ensemble classification, in order to have the maximum amount of data clustered in one of the identified groups. The idea is to now see this as a classification problem, i.e. using the clustered data as the training set to train the classifiers and use the unclustered data as the test set.

An abstract view for the steps of the novel pipeline is shown in Figure 1. The first step of the pipeline is data gathering, followed by cleaning in the preprocessing unit. The processed data are passed to the 'Coupling ensembles of Clustering and classification' unit which is itself broken down in further substeps, each discussed in the next section. The final clustering in groups is inspected in the Further analysis step. The units of the pipeline are explained in the following subsections.

3.1. Data Gathering

The first step in the pipeline is aggregation of the data. Data can come in different shapes and formats, but the process and protocols to collect the data are out of scope of this paper.

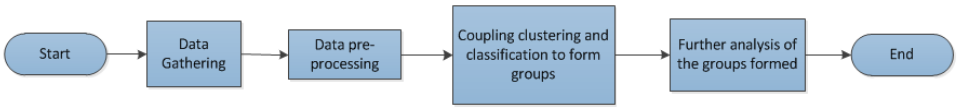


Figure 1: Abstract flowchart of the steps in the proposed pipeline

3.2. Data Pre-processing

The next step is data pre-processing, i.e. data cleaning, which enhances the power of algorithms to identify the patterns inside the datasets. In order to fully understand the data under inspection descriptive statistics such as mean, median, range, standard deviation, etc. can be computed. This might also help to locate inconsistencies among the data. Tuples which have missing values can be deleted, and the tuples attributes could be made homogeneous, i.e. all numeric or text. Other ways for processing could be normalisation of values using methods such as min-max, z-score, etc. [45]. These steps produce the final set of data points to be used.

3.3. Coupling ensembles of Clustering and Classification

Coupling ensembles of Clustering and Classification is composed of multiple sub-steps, as depicted by a detailed flowchart in Figure 2. In this unit, the first sub-step is Ensemble Clustering Unit, explained in the following subsection.

3.3.1. Ensemble Clustering (Clusterfusion) Unit

This section elucidates the construction of groups and initial data distribution from the dataset by consensus among the results of the two clustering algorithms. The cluster ensemble methodology looks similar in principle to equivalent approaches in classification, namely ensemble methods like bagging and boosting, but it presents additional problems [29]. Firstly, matching clusters between algorithms is not straightforward as different clustering algorithms may generate different numbers of groups and moreover the optimal number of groups may be unknown. Another common problem is that of unclustered data left while combining the results of the clustering algorithms.

To determine the optimal value of number of clusters 'k', we suggest to run cluster validity indices. Set the minimum and maximum values for k, then at each iteration run the clustering algorithms and the validity indices, changing the input k between the two set values. If the indices suggest different numbers of clusters, then the minimum sum of ranks can be taken as the optimal value of k [2]. This information about k is now passed back to cluster the data using the same two clustering algorithms. To verify the ensemble solution, Swift et al. [46] suggested the use of the kappa index. The agreement between the equivalent clusters returned by the different clustering algorithms is evaluated using the Cohen's unweighted and weighted kappa index [47]. The index statistically evaluates the agreement between the two clustering algorithms which clusters n data among k groups, and return a value between 0-1 with zero being absolute disagreement and one being complete agreement. Thus, this process returns the most stable cluster size.

The groups obtained by the different clustering algorithms are merged with the goal of assigning the same set of objects to identical clusters, a process known as 'correspondence problem'. The correspondence among the groups is achieved by aligning the group labels to get maximum agreement and therefore form core groups composed of the data with the same group label. These steps of the ensemble clustering unit are represented by a flowchart in Figure 3. There might be some data that present mixed classification while dividing the data points into k common groups. These remaining data are therefore called unclustered

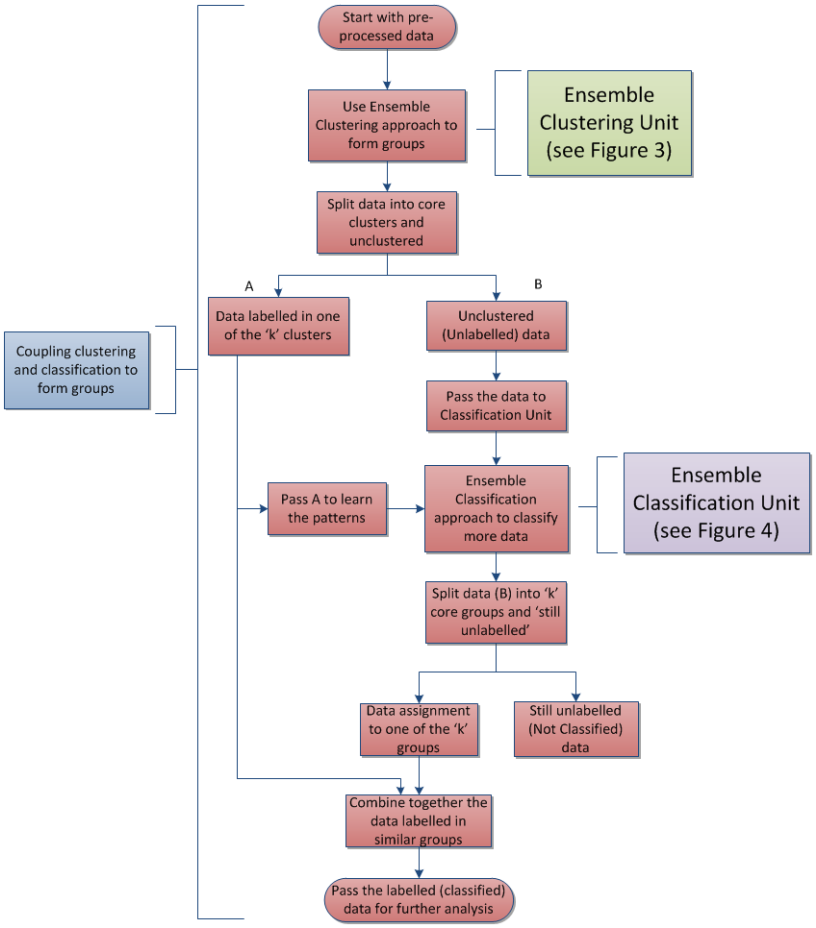


Figure 2: Flowchart of the sub-steps in 'Coupling Clustering and Classification to form groups' step from the abstract pipeline view. 'Ensemble clustering (Clusterfusion) to form groups' and 'Ensemble Classification approach to classify more data' are further subdivided into smaller steps explained in Figures 3 and 4 respectively.

(unlabelled). To tackle the unclustered data we introduce the use of Ensemble Classification on top of clusterfusion and forward the unclustered data to the 'Ensemble Classification Unit'.

3.3.2. Ensemble Classification Unit

The ensemble classification unit is introduced to refine the clustering results by targeting the data that present mixed classification. The reasons behind introducing the ensemble classification in the pipeline are to improve solutions, to reuse knowledge and to select novel models by changing classifiers. Both the classification algorithms first learn the classification rules from the previously clustered data (training set), and then use the unclustered one as the test set with the goal to assign them to one of the previously identified 'k' groups. Decision-level fusion using unanimous voting is employed to merge the clas-

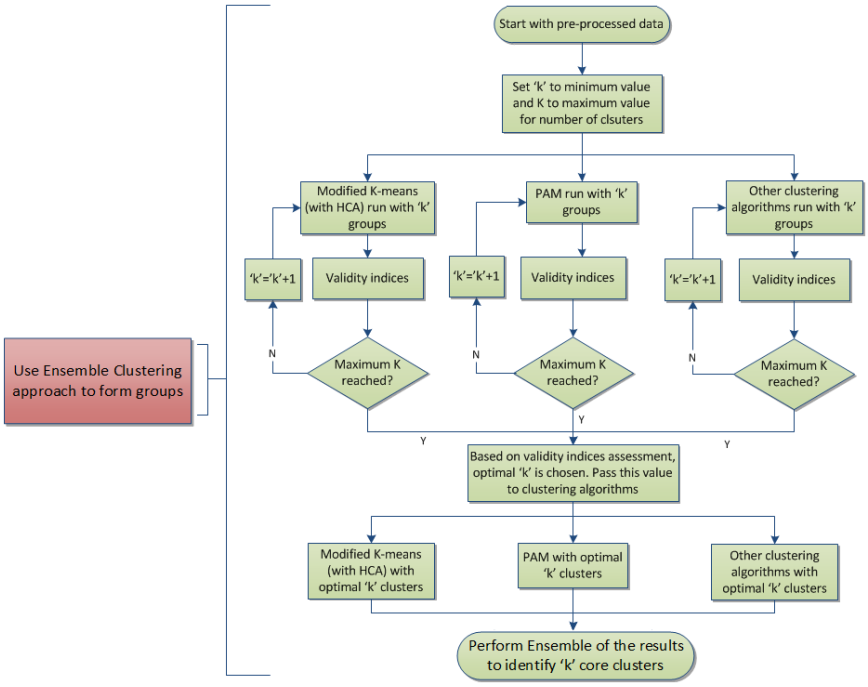


Figure 3: Flowchart for Consensus Clustering Unit shows the pictorial representation of the order of steps to form groups in a dataset and subsequently divide the data in these constitued groups.

sification results. The reason for using this method is the choice of real world application to assign breast cancer patients to biological classes, which demands high confidence in the results. As Unanimous voting considers the consensus across all the algorithms in the ensemble, it generates the highest confidence in results among all the ensemble methods.

Two principles were considered while performing the consensus: i) include as many classifiers in the ensemble as possible, and ii) assign as many data to the core groups as possible. Strict application of the principles may result in few data assignments, leading to conflicts between them. The possible solution for this problem can be: i) to use other methods of ensemble classifiers such as majority voting, weighted voting, stacked generalisation, etc. ii) to use a trade off between the two principles. However, this problem is currently overlooked as high output confidence is desired in this work.

After ensemble classification, more data are assigned to one of the groups, and the remaining data are still unlabelled or 'Not classified'. The data clustered after the clusterfusion step are now combined with the newly clustered data in the same groups to achieve the final clusters. The newly clustered data needs further analysis and verification, which are discussed in the following section. The detailed flowchart for the order of steps in this unit is shown in Figure 4.

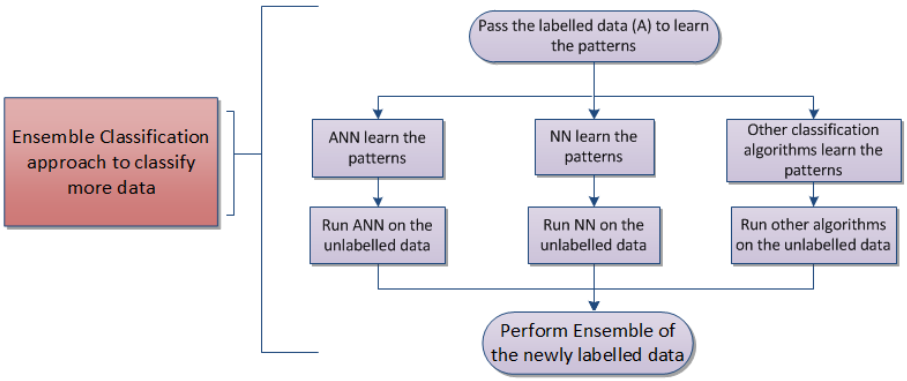


Figure 4: Flowchart for the order of steps in Consensus Classification Unit to classify more data in groups constituted in the clusterfusion step.

3.4. Further Analysis of the Groups Formed

To evaluate and validate the characteristics of the data assigned after ensemble classification, visual tests, statistical tests and cluster quality assessment need to be performed. The distribution of each attribute in a group after clusterfusion was compared to the one after ensemble classification using boxplots (a method of visually displaying the distribution of the data using median, quartiles and outliers [48]) and the Davies-Bound index (internal cluster quality assessment index). The statistical test used for the comparison is Mann-Whitney-Wilcoxon, a non-parametric version of the t-test, to check whether the two distributions come from the same population. The level of significance is usually chosen at 0.05 [49].

4. Results

This section show the results for the application of the pipeline on the real world Breast Cancer Dataset and Standard Benchmark Datasets.

4.1. Real World Dataset: Edinburgh Breast Cancer Series

The real world data used was a breast cancer dataset provided by the Edinburgh Research Centre. The Edinburgh Breast Cancer Series consists of 885 patients treated with breast conservation surgery, axillary node sampling or clearance, and whole breast radiotherapy [6]. Each of the patient's tumour samples were tested against ten biomarkers using immunohistochemistry, which resulted in a score for each biomarker ranging from 0 to 300. The biomarkers were: Estrogen Receptor (ER), Progesterone Receptor (PgR), Cytokeratin 7/8 (CK7/8), Cytokeratin 5/6 (CK5/6), EGFR, c-erbB2 (HER2), c-erbB3 (HER3), c-erbB4 (HER4), p53 and Mucin1 (MUC1) [3], [4].

The next step after selecting the dataset is data processing. In the Edinburgh breast cancer dataset, the biomarkers ER, PgR and HER2 had no missing values for all the patients while other biomarkers had few missing entries. For the purpose of our pipeline we deleted all the tuples with at least one missing value,

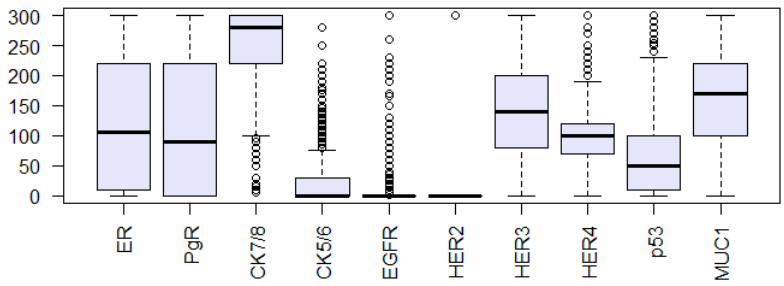


Figure 5: Boxplots of 478 patients in the Edinburgh Dataset

which reduced the number of patients down to 478. Since all the biomarkers were scored on a 0 to 300 range, normalisation was not required. The cleaning of the cancer dataset was followed by the computation of statistical measures (namely median, range and outliers) and boxplots. Figure 5 shows the boxplots of the 478 breast cancer patients where the y-axis represents the values for each biomarker and the x-axis represents the ten biomarkers.

After the descriptive analysis, the next step is to choose the number of groups by employing PAM and K-means clustering algorithms for 'k' = 2 to 15. The cluster results obtained were passed to the cluster validity indices to asses the quality of the clustering output and hence determine the optimal number of groups. Each validity index (Marriot, Calinski, Scott and TraceW) was run separately on each clustering algorithm, with an associated rule to choose the best number of clusters, already shown in Table 1. All the Cluster Validity indices suggested four to be the optimal number except TraceW when run after K-means, for which the optimal number was three. The results of the validity indices computation for both the algorithms are shown in Table 2. The minimum sum of ranks was taken and four was chosen as the optimal number.

Both the clustering algorithms were again run on the dataset for four groups. The correspondence between results of the two clustering algorithms were statistically compared using the unweighted and the weighted Cohen's kappa index. The weights of the weighted kappa index were computed based on the degree of disagreement between the groups, i.e. the higher the weight, higher the disagreement. The unweighted kappa index value was 0.85 and the weighted kappa index value was 0.93. These two values indicate a very high agreement between the two clustering algorithms for the four groups formed. The cluster labels of both the clustering algorithms were aligned in order to have the same data assigned to the cluster with the same label from the two methods. The distribu-

Validity Index	PAM	K-means
Marriot	4	4
Calinski	4	4
Scott	4	4
TraceW	4	3

Table 2: Number of clusters suggested by different validity indices

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Unclustered
K-means	168	141	120	49	-
PAM	160	190	81	47	-
Ensemble Clustering (Clusterfusion)	157	141	81	47	52
Ensemble Classification (additionally assigned)	13	12	11	2	14
Coupling Cluterfusion with ensemble classification	170	153	92	49	14

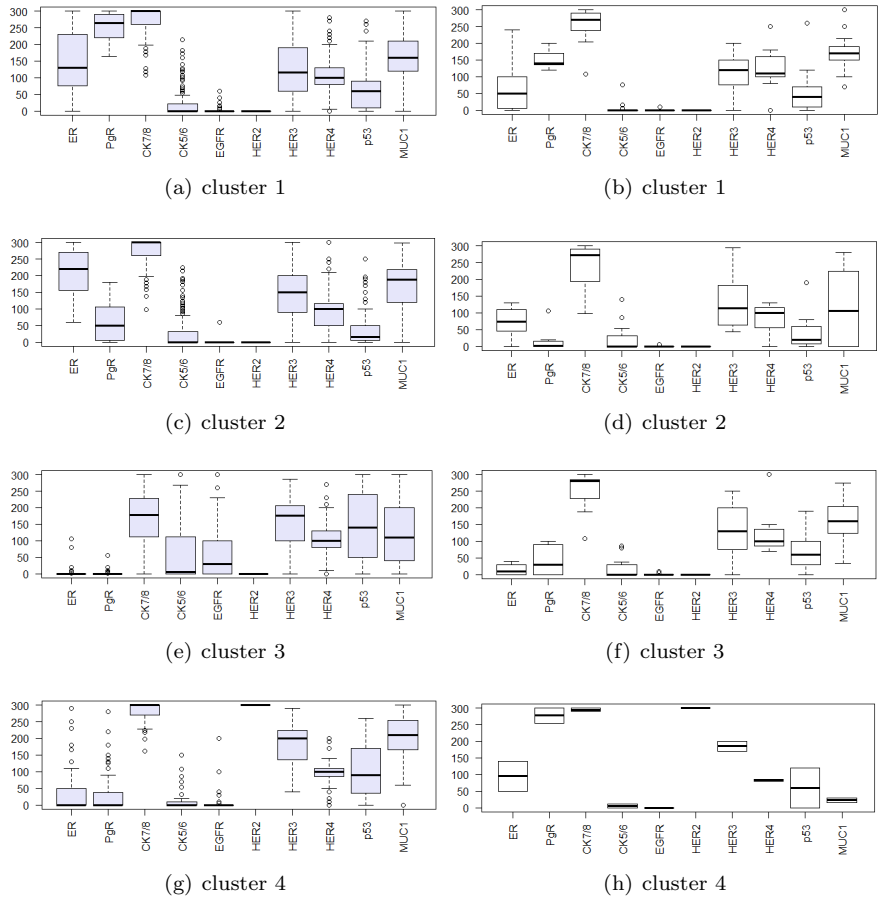
Table 3: Division of patients in clusters

tion of patients among the four classes is shown in the first two rows of Table 3. Clusters 1, 2, 3 and 4 represent the four identified groups. Clusterfusion grouped 426 patients among the four common clusters and left 52 unclustered (row three of Table 3).

The next step was to train the two classifiers on the 426 clustered patients, run them on the 52 unclustered patients as test set and perform the class level fusion of results. The first classifier used was five layered ANN network (having three hidden layers), with a sigmoid activation function to model the weighted sum. The ANN was tested with multiple models, but it achieved best results for network having 7-10-7 nodes for the hidden layers, with 200 epochs. The second classifier was the Nearest Neighbour with an Euclidean distance function. For both ANN and Nearest Neighbour, we used a 10 fold cross validation approach. An additional 38 patients were assigned to the common groups whose distribution is shown in row four of Table 3. Combining clusterfusion and ensemble classification, a total of 464 patients were clustered in one of the groups and the final distribution among the clusters is shown in the last row of Table 3.

The data assigned to groups after ensemble classification were visually verified using boxplots. Figure 6 compares the boxplots of the biomarkers for patients clustered in four groups after clusterfusion with the boxplots for the patients subsequently assigned to the groups among the unclustered ones. The boxplots on the left side, i.e. (a),(c),(e) and (g), are of 426 patients after clusterfusion in clusters 1, 2, 3 and 4 respectively. Similarly, boxplots on the right side, i.e. (b),(d),(f) and (h), are of 38 more patients assigned after ensemble classification in clusters 1, 2, 3 and 4 respectively. Of these 38, 13 were assigned to cluster 1, 12 to cluster 2, 11 to cluster 3 and two to cluster 4.

It can be observed that the boxplots of the patients after clusterfusion are very similar to the ones after ensemble classification for corresponding groups implying differences between the medians may not be significant. Few biomarkers differ, as in cluster 1 after clusterfusion, the medians of ER and PgR are in the range of 100 to 150 and 250 to 300 respectively, while after ensemble classification the median is around 50 for ER and 150 for PgR. Cluster 2 shows the variation only for one biomarker, ER, with its median in the range of 200 to 250



(a) cluster 1 (b) cluster 1

(c) cluster 2 (d) cluster 2

(e) cluster 3 (f) cluster 3

(g) cluster 4 (h) cluster 4

Figure 6: Boxplots of biomarkers. (a), (c), (e), (g): for patients in clusters derived after clusterfusion. (b), (d), (f), (h): for additional patients classified after ensemble classification.

for patients after clusterfusion and being in the range 50 to 100 after ensemble classification. Cluster 3 also shows a difference for only one biomarker, CK7/8, with the median being in the range of 150 to 200 for patients after clusterfusion and in the range of 250 to 300 for patients after ensemble classification. The differences among biomarkers in cluster 4 are among ER, PgR and MUC1: the medians after clusterfusion are almost 0 for ER and PgR and just more than 200 for MUC1, while for patients clustered after ensemble classification the value of the median for ER is almost 100, between 250-300 for PgR, and that of MUC1 is below 50. For other biomarkers in all the classes, the medians are almost equal. The reported differences in the values of the biomarkers could be the reason that patients grouped after ensemble classification were originally unclustered after clusterfusion.

We now check whether the aforementioned differences are statistically significant, for which we compare the distributions of markers in the newly formed groups with the groups from clusterfusion, using the Mann-Whitney-Wilcoxon

Biomarker	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ER	0.004	$2.34\text{x}10^{-7}$	0.125	$8.02\text{x}10^{-6}$
PgR	$5.69\text{x}10^{-9}$	0.003	0.001	$6.89\text{x}10^{-5}$
CK7/8	0.061	0.072	0.005	0.681
CK5/6	0.158	0.539	0.214	0.84
EGFR	0.937	0.988	0.0004	0.467
HER2*	1	1	1	1
HER3	0.587	0.074	0.356	0.879
HER4	0.25	0.962	0.751	0.202
p53	0.438	0.766	0.0325	0.39
MUC1	0.637	0.055	0.082	0.024

*The reason for the p-value for HER2 to be equal to 1 is because the HER2 H-score for all patients in the data was either 0 (negative) or 300 (positive), and it did not follow exactly a continuous range of values as for the other biomarkers.

Table 4: The p-value comparison for patients after clusterfusion with ensemble classification

test. Table 4 shows the p-values comparison of 426 patients after clusterfusion with 38 after ensemble classification for all the clusters and markers. The level of significance was set at 0.05. In both clusters 1 and 2, it can be seen that ER and PgR have statistically different distributions. In cluster 3, significant differences are observed for PgR, CK7/8, EGFR and p53. In cluster 4, the biomarkers ER, PgR and MUC1 are significantly different, similar to the boxplots observation. Based on the boxplots and statistical results, the new patients added after ensemble classification could be assigned to one of the groups.

To understand the significance of adding the ensemble classification layer on top of the clusterfusion layer, the Davies Bound (DB) [21] internal cluster validity index is used, which is one of the most stable and commonly used indexes in literature. The DB index computes the ratio of intra- and inter-cluster distance of the points (i.e. patients), therefore, the lower this value the better the clustering. Since additional patients are added to the clusters, the value of DB will slightly increase. As the value of the DB index increases, the quality of the final classification decreases and thus the lower the increase, the better the classification. Therefore, to study the quality of clusters, the DB index values of each subgroup of patients assigned after ensemble classification were compared to the values of the DB index obtained if they had been assigned to each of the other clusters.

The DB index value of the overall classification after clusterfusion is 1.3732. Table 5 shows the DB index value of each subgroup of patients assigned to each cluster. Each column represents the cluster predicted for a set of patients after ensemble classification, i.e. third row of Table 3. Each row contains the DB index of the assigned cluster. It can be observed from the table that for Clusters 1, 2, and 3, ensemble classification assigns patients to the best possible group, while for cluster 4, the two additional patients were assigned to the second best group. Overall, the DB index supports the distribution of patients after ensemble classification. It was also observed that at least one of the clustering algorithms did assign the elements to the same class as obtained after ensemble classification. Therefore, the use of the DB index supports this, and the clusters

Assigned Cluster	13 patients from Cluster 1	12 patients from Cluster 2	11 patients from Cluster 3	2 patients from Cluster 4
Cluster 1	**1.3840**	1.4004	1.3933	**1.3736**
Cluster 2	1.4339	**1.3877**	1.4297	1.3835
Cluster 3	1.4138	1.3991	**1.3915**	1.3893
Cluster 4	1.4016	1.4442	1.3941	**1.3749**

Table 5: DB index value of each subgroup of patients assigned after ensemble classification, compared by assigning to all the other remaining clusters.

in which the patients are added after ensemble classification is the best possible.

Clinical Assessment

We can be confident in assigning the patients after ensemble classification to the biological classes identified after clusterfusion. The final clusters (after combining clusterfusion and ensemble classification) represent the four biological groups reported in Figure 7, shown using a tree format. According to available pathological information ER and PgR levels define Luminal groups. The two groups are characterised by positive levels for ER values but different PgR levels, i.e. positive and negative respectively. Thus, Clusters 1 and 2 represent Luminal A and Luminal B biological groups respectively. Regarding those patients assigned to groups after ensemble classification, although values may differ in some biomarkers, the patterns of the defining biomarkers need to be consistent. In Clusters 1 and 2 patients after ensemble classification follow the same trend, i.e. ER positive/PgR positive and ER positive/PgR negative respectively.

Cluster 3 represents Basal class with p53 altered, characterised by high levels of p53 and low levels of ER. Patients in cluster 3 after ensemble classification have high levels of p53 and low levels of ER similar to the patients after clusterfusion. Cluster 4 characterise HER2+ tumours, having high values of HER2 biomarkers. The patients after clusterfusion have low levels of biomarker ER, called HER2/ER- group. However, the patients assigned to cluster 4 after ensemble classification have high levels of ER (called HER2/ER+), which could be the reason for them being unclustered. Since the number of patients after ensemble classification are very small i.e. 2, thus they are labelled together under one biological class called HER2+.

4.2. Benchmark Datasets

Five UCI benchmark datasets [24] including two standard medical datasets were used to evaluate the framework i.e. external validation. These five datasets were selected as they contain different number of classes with a broad range of features, shown in Table 6. The pipeline was run on all the datasets by removing the ground truth from them. After distributing the data among the groups a comparison with the ground truth of each dataset was conducted.

Table 7 shows the results of the pipeline on the UCI datasets. The first row contains the number of data clustered in the groups after ensemble clustering represented by 'Number of data clustered after Ensemble Clustering'. The second row contains the number of data correctly grouped after ensemble clustering by comparing with the ground truth represented by 'Number of data correctly clustered after Ensemble Clustering'. The third row contains the number of

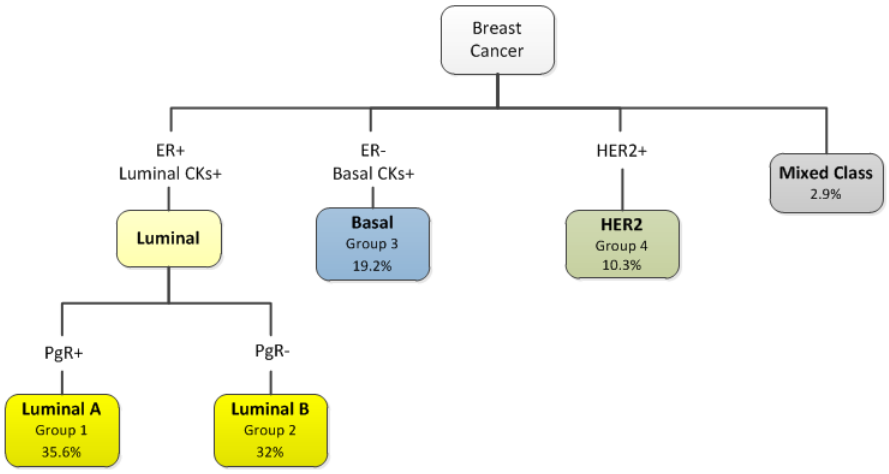


Figure 7: Breast Cancer Groups with representative characteristics

	Iris	Wine	Ecoli	Dermatology	WBC*
Number of Classes	3	3	8	6	2
Number of Features	4	13	7	33	9
Total Number of Data	150	178	336	366	683

*Wisconsin Breast Cancer

Table 6: Descriptive features such as number of classes, number of features and total data in the benchmark datasets

data clustered in the groups after ensemble classification and the fourth row the number of data correctly grouped after ensemble classification represented by 'Number of data clustered after Ensemble Classification' and 'Number of data correctly clustered after Ensemble Classification', respectively. 'Total data correctly clustered' contains the sum of number of data correctly clustered after ensemble clustering and ensemble classification. The last two rows contain the percentage of data correctly clustered after ensemble clustering and the percentage of total data accurately clustered respectively. The increase in percentage shows the improvement in results upon adding ensemble classification stage, as discussed in the next section.

5. Discussion

While exploring the literature for the division of patients in biological classes using ensemble clustering, the need for improving the clustering results was identified. For example, Soria et al. [2] used clusterfusion on a breast cancer dataset but could only classify 62% of the patients into one of the groups. Kellam et al. [13] also used clusterfusion (on a different domain though) and could only classify 55% of the patients. Thus, in this work we introduce the use of ensemble classification to improve the distribution by targeting the unclustered data and present an improved pipeline for the identification of groups and data

	Iris	Wine	Ecoli	Dermatology	WBC*
Number of data clustered after Ensemble Clustering	134	170	334	351	681
Number of data correctly clustered after Ensemble Clustering	133	161	289	335	655
Number of data clustered after Ensemble Classification	16	8	2	15	2
Number of data correctly clustered after Ensemble Classification	11	7	2	15	2
Total Data correctly clustered after combining Ensemble Clustering and Ensemble Classification	144	168	291	350	657
Percentage of Data correctly clustered after Ensemble Clustering	88.67%	90.5%	86%	91.5%	95.9%
Percentage of Total Data correctly clustered after combining Ensemble Clustering and Ensemble Classification	96%	94.4%	86.6%	95.6%	96.2%

*Wisconsin Breast Cancer

Table 7: A comparison of distribution of data among the groups for the benchmark datasets.

distribution in them for an unlabelled dataset. K-means (with HCA) and PAM clustering algorithms have been used for clusterfusion as they are two of the most commonly used clustering algorithms with good performance [50]. Subsequently, an ensemble classification was used to distribute more patients, which were previously unclustered. Two classifiers (ANN and the Nearest Neighbour) have been utilised for the ensemble classification phase [51], as together they have consistently resulted in good performance throughout the literature [52] and cover advantageous properties required for good ensemble classification such as non-linearity, non-parametric, memory based learning, etc. The application of the pipeline on the Edinburgh Series identified four possible biological classes and distributed the patients among them. Using ensemble clustering 89% of the patients were assigned to one of the groups, but after adding ensemble classification the classification jumped to 97%.

Additionally, Soria and colleagues in two consequent works ([2] and [3]) proved the existence of three major biological classes: Luminal, Basal and HER2. Therefore, by comparing the classification results of Edinburgh dataset with the available pathological information (and based on the characteristics), we have obtained similar biological groups as in the literature. The patients were distributed among four biological groups using the pipeline, as cluster 1 can be labelled as Luminal A, cluster 2 as Luminal B, cluster 3 as Basal, cluster 4 as HER2+ and the remaining patients as Mixed Class (Figure 7).

To measure the effectiveness, robustness and generalisation of the proposed framework, it was also run on standard datasets by removing their ground truth. A significant improvement in the classification results was also observed in all the standard datasets. Experimental results show that the proposed framework effectively finds groups in the datasets, distributes the data points in them and finally validates the results with the ground truth. Although the pipeline does

not achieve the highest accuracy on standard datasets in comparison to other supervised classification algorithm accuracies in literature, these differences are small and acceptable. Moreover, these differences occur as the pipeline takes the unlabelled data as input while the supervised algorithms learn the classification rules on the ground truth in the training set. Thus, a trade-off can be considered to identify groups in the datasets with no class labels and to get most of the data correctly clustered.

As discussed, one limitation of the proposed pipeline could be overfitting. We tried to address this issue by testing the framework and running experiments on standard datasets with ground truth. Another limitation could be its specific settings for the Edinburgh dataset, and the pipeline settings need to be adjusted for other datasets. Finally, the third limitation of the work might be the use of ensemble method as it is more restrictive and less flexible, thus reducing the number of data assigned to the groups. To address this limitation, we tested our pipeline on the unclustered data left behind by Soria et al. [3], who have made use of clusterfusion and fuzzy rule based algorithms to identify and assign data in groups on a breast cancer dataset collected at the Nottingham City Hospital [53]. They divided 1,035 breast cancer patients out of 1,073 among seven biological classes. Using our pipeline, 18 more patients were assigned to one of the biological groups [19], verified by clinical experts.

6. Conclusions

In the domain of breast cancer, the identification of biological diversity is extremely important for clinical experts to determine the best treatment outcome. Previously, ensemble clustering algorithms have been used to elucidate core biological classes in a breast cancer dataset. However, this methodology results in a number of unclustered patients, i.e. low classification in biological groups. Thus, in this paper we introduced the use of ensemble classification and presented a novel pipeline for improved data distribution in the identified groups. An ensemble of multiple clustering algorithms has been used for the extraction of characteristic biological classes from a breast cancer dataset and initial distribution of patients in them. The final data distribution in groups was achieved by combining the patients distributed after ensemble clustering with the patients distributed after ensemble classification.

The application of this pipeline on a real world breast cancer dataset is presented. The kappa index of agreement between the outcomes of the two clustering algorithms was high, indicating a good agreement between the two techniques. Initially, after Ensemble Clustering, 426 data out of 478 were clustered among one of the four groups. After adding the ensemble classification unit to the methodology, a total of 464 patients were clustered. Mann-Whitney-Wilcoxson test and boxplots suggested the data grouped after ensemble classification were very similar to the data grouped after clusterfusion. We have subsequently used the DB index to test the quality of the clusters, to verify that after ensemble classification patients were added to the most appropriate cluster. The characteristics of the obtained biological groups were similar to the corresponding ones in the literature.

We have also presented the application of this novel pipeline on several standard datasets. The distribution of the data in the groups was verified with already present ground truth for them. Although the final accuracy was equivalent to what is achieved by other classifiers in the literature on the supervised algorithms, these difference are acceptable. The results on the standard dataset show the effectiveness of the pipeline, particularly for those datasets without labels.

This work expands the initial idea of employing the ensemble classification method in a breast cancer identification pipeline [19], and its focus is not on the number of clusters formed but on the overall methodology to refine clustering results. The biggest advantage of the proposed pipeline lies in its adaptability, which enables to customise each component according to need, and at each stage. Although in this paper we have used specific clustering algorithms, classification algorithms, ensemble methods and cluster validity indices, all of them could be replaced by similar ones. The results also prove the superiority of using an ensemble approach as the groups obtained are more robust. The ensemble not only provides robustness to the groups, but to the methodology as well.

In the future, we plan to use other class level fusion methods to assess the results and increase the number of classifiers in the ensemble. We also plan to apply the methodology on different datasets (not limited to medical purposes) and test its proficiency. The proposed pipeline has shown promising results in refining and improving the number of data clustered, but further validation will be sought.

7. Acknowledgements

The research was supported by The University of Nottingham Vice-Chancellor's Scholarship for Research Excellence (International). The authors would like to thank Javier Navarro Barron for his valuable comments.

References

[1] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (1) (2013) 243–256. doi:10.1016/j.patcog.2012.07.021.

[2] D. Soria, J. M. Garibaldi, F. Ambrogi, A. R. Green, D. Powe, E. Rakha, R. Douglas Macmillan, R. W. Blamey, G. Ball, P. J. G. Lisboa, T. A. Etchells, P. Boracchi, E. Biganzoli, I. O. Ellis, 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. doi:10.1016/j.compbiomed.2010.01.003.

[3] D. Soria, J. M. Garibaldi, a. R. Green, D. G. Powe, C. C. Nolan, C. Lemetre, G. R. Ball, I. O. Ellis, A quantifier-based fuzzy classification system for breast cancer patients, in: Artificial Intelligence in Medicine, Vol. 58, 2013, pp. 175–184. doi:http://dx.doi.org/10.1016/j.artmed.2013.04.006.

[4] E. A. Rakha, D. Soria, A. R. Green, C. Lemetre, D. G. Powe, C. C. Nolan, J. M. Garibaldi, G. Ball, I. O. Ellis, Nottingham Prognostic Index Plus (NPI+): a modern clinical decision making tool in breast cancer., British journal of cancer 110 (7) (2014) 1688–97. doi:10.1038/bjc.2014.120.

[5] D. Soria, J. M. Garibaldi, A novel framework to elucidate core classes in a dataset, 2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 IEEE Congress on Evolutionary Computation, CEC 2010arXiv:arXiv:1011.1669v3, doi:10.1109/CEC.2010.5586331.

[6] A. R. Green, D. Soria, J. Stephen, D. G. Powe, C. C. Nolan, I. Kunkler, J. Thomas, G. R. Kerr, W. Jack, D. Cameron, T. Piper, G. R. Ball, J. M. Garibaldi, E. A. Rakha, J. M. Bartlett, I. O. Ellis, Nottingham Prognostic Index Plus: Validation of a clinical decision making tool in breast cancer in an independent series, The Journal of Pathology: Clinical Research 2 (1) (2016) 32–40. doi:10.1002/cjp2.32.

[7] L. Rokach, O. Maimon, Chapter 15 Clustering methods, The Data Mining and Knowledge Discovery Handbook (2010) 32doi:10.1007/0-387-25465-X_15.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns., Proceedings of the AMIA Symposium. American Medical Informatics Association 95 (25) (1998) 14863–8. doi:10.1073/pnas.95.25.14863.

[9] S. E. Elsheikh, A. R. Green, E. A. Rakha, D. G. Powe, R. A. Ahmed, H. M. Collins, D. Soria, J. M. Garibaldi, C. E. Paish, A. A. Ammar, M. J. Grainge, G. R. Ball, M. K. Abdelghany, L. Martinez-Pomares, D. M. Heery, I. O. Ellis, Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome, Cancer Research 69 (9) (2009) 3802–3809. doi:10.1158/0008-5472.CAN-08-3907.

[10] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. a. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. a. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, a. L. Børresen-Dale, P. O. Brown, D. Botstein, Molecular portraits of human breast tumours., Nature 406 (6797) (2000) 747–752. arXiv:NIHMS150003, doi:10.1038/35021093.

[11] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, A. L. Børresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications., Proceedings of the National Academy of Sciences of the United States of America 98 (19) (2001) 10869–74. doi:10.1073/pnas.191367098.

[12] C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, E. T. Liu, Breast cancer classification and prognosis based on gene expression profiles from a population-based study, Proc Natl Acad Sci U S A 100 (18) (2003) 10393–10398. doi:10.1073/pnas.1732912100.

[13] P. Kellam, X. Liu, N. Martin, C. Orengo, S. Swift, A. Tucker, Comparing, Contrasting and Combining Clusters in Viral Gene Expression, Proceedings of the Sixth Workshop on Intelligent Data Analysis in Medicine and Pharmacology.

[14] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, Machine Learning 52 (1-2) (2003) 91–118. doi:10.1023/A:1023949509487.

[15] D. Chen, K. Xing, D. Henson, L. Sheng, A. M. Schwartz, X. Cheng, Developing prognostic systems of cancer patients by ensemble clustering., Journal of biomedicine & biotechnology 2009. doi:10.1155/2009/632786.

[16] D. Soria, J. M. Garibaldi, E. Biganzoli, I. O. Ellis, A Comparison of Three Different Methods for Classification of Breast Cancer Data, 2008 Seventh International Conference on Machine Learning and Applications (2008) 619–624doi:10.1109/ICMLA.2008.97.

[17] A. R. Green, D. G. Powe, E. A. Rakha, D. Soria, C. Lemetre, C. C. Nolan, F. F. T. Barros, R. D. Macmillan, J. M. Garibaldi, G. R. Ball, I. O. Ellis, Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers., British journal of cancer 109 (7) (2013) 1886–94. doi:10.1038/bjc.2013.528.

[18] D. Vuong, P. T. Simpson, B. Green, M. C. Cummings, S. R. Lakhani, Molecular classification of breast cancer, Virchows Archiv 465 (1) (2014) 1–14. doi:10.1007/s00428-014-1593-7.

[19] U. Agrawal, D. Soria, C. Wagner, Cancer Subtype Identification Pipeline: A Classifusion Approach, IEEE World Congress on Computational Intelligence, WCCI 2016 - IEEE Congress on Evolutionary Computation, CEC 2016.

[20] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 42 (4) (2012) 463–484. doi:10.1109/TSMCC.2011.2161285.

[21] E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz, Internal versus External cluster validation indexes, International Journal of Computers and Communications 5 (1) (2011) 27—-34.

[22] T. Chakraborty, Ec3: Combining clustering and classification for ensemble learning, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 781–786.

[23] J. Zhang, X. Chen, Y. Xiang, W. Zhou, J. Wu, Robust network traffic classification, IEEE/ACM Transactions on Networking (TON) 23 (4) (2015) 1257–1270.

[24] M. Lichman, {UCI} Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2013.

[25] G. Tzortzis, A. Likas, The MinMax k-Means clustering algorithm, Pattern Recognition 47 (7) (2014) 2505–2516. doi:10.1016/j.patcog.2014.01.015.

[26] M. E. Celebi, H. A. Kingravi, P. A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications 40 (1) (2013) 200–210. arXiv:arXiv:1209.1960v1, doi:10.1016/j.eswa.2012.07.021.

[27] H. S. Park, C. H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications 36 (2 PART 2) (2009) 3336–3341. doi:10.1016/j.eswa.2008.01.039.

[28] A. Strehl, J. Ghosh, Cluster Ensembles A Knowledge Reuse Framework for Combining Multiple Partitions, Journal of Machine Learning Research 3 (2002) 583–617. doi:10.1162/153244303321897735.

[29] J. Ghosh, A. Acharya, Cluster ensembles, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (4) (2011) 305–315. doi:10.1002/widm.32.

[30] L. Rokach, Ensemble-based classifiers, Artif Intell Rev 33 (2010) 1–39. doi:10.1007/s10462-009-9124-7.

[31] S. VEGA-PONS, J. RUIZ-SHULCLOPER, A Survey of Clustering Ensemble Algorithms, International Journal of Pattern Recognition and Artificial Intelligence 25 (03) (2011) 337–372. doi:10.1142/S0218001411008683.

[32] F. Amato, A. López, E. M. Peña-Méndez, P. Vahara, A. Hampl, J. Havel, Artificial neural networks in medical diagnosis, Journal of Applied Biomedicine 11 (2) (2013) 47–58. doi:10.2478/v10136-012-0031-x.

[33] A. Pouliakis, E. Karakitsou, N. Margari, P. Bountris, M. Haritou, J. Panayiotides, D. Koutsouris, P. Karakitsos, Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future., Biomedical engineering and computational biology 7 (2016) 1–18. doi:10.4137/BECB.S31601.

[34] A. Mert, N. Klç, E. Bilgili, A. Akan, Breast Cancer Detection with Reduced Feature Set, Computational and Mathematical Methods in Medicine 2015 (2015) 1–11. doi:10.1155/2015/265138.

[35] M. Sarkar, T. Y. Leong, Application of K-nearest neighbors algorithm on breast cancer diagnosis problem., Proceedings of the AMIA Annual Symposium. American Medical Informatics Association (2000) 759–763.

[36] S. A. Medjahed, T. A. Saadi, A. Benyettou, Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules, International Journal of Computer Applications 62 (1) (2013) 975–8887. doi:10.5120/10041-4635.

[37] S. Oh, M. S. Lee, B. T. Zhang, Ensemble learning with active example selection for imbalanced biomedical data classification, IEEE/ACM Transactions on Computational Biology and Bioinformatics 8 (2) (2011) 316–325. doi:10.1109/TCBB.2010.96.

[38] A. Acharya, E. R. Hruschka, J. Ghosh, S. Acharyya, C3E: A framework for combining ensembles of classifiers and clusterers, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6713 LNCS (2011) 269–278. doi:10.1007/978-3-642-21557-5_29.

[39] D. M. Farid, L. Zhang, A. Hossain, C. M. Rahman, R. Strachan, G. Sexton, K. Dahal, An adaptive ensemble classifier for mining concept drifting data streams, Expert Systems with Applications 40 (15) (2013) 5895–5906. doi:10.1016/j.eswa.2013.05.001.

[40] R. Lysiak, M. Kurzynski, T. Woloszynski, Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers, Neurocomputing 126 (2014) 29–35. doi:10.1016/j.neucom.2013.01.052.

[41] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, D. Amorim Fernández-Delgado, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, Journal of Machine Learning Research 15 (2014) 3133–3181. doi:10.1016/j.csda.2008.10.033.

[42] P. Domingos, A few useful things to know about machine learning, Communications of the ACM 55 (10) (2012) 78. arXiv:9605103, doi:10.1145/2347736.2347755.

[43] S. Bashir, U. Qamar, F. H. Khan, Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble, Quality and Quantity 49 (5) (2014) 2061–2076. doi:10.1007/s11135-014-0090-z.

[44] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woźniak, Ensemble learning for data stream analysis: a survey, Information Fusion 37 (2017) 132–156. doi:10.1016/j.inffus.2017.02.004.

[45] S. G. K. Patro, K. K. Sahu, Normalization: A Preprocessing Stage, arXiv preprint arXiv:1503.06462. (2015) 4arXiv:1503.06462.

[46] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, P. Kellam, Consensus clustering and functional interpretation of gene-expression data., Genome biology 5 (11) (2004) R94. doi:10.1186/gb-2004-5-11-r94.

[47] Z. Yang, M. Zhou, Kappa statistic for clustered matched-pair data, Statistics in Medicine 33 (15) (2014) 2612–2633. doi:10.1002/sim.6113.

[48] M. Krzywinski, N. Altman, Visualizing samples with box plots, Nature Publishing Group 11 (2). doi:10.1038/nmeth.2813.

[49] T. Sellke, M. J. Bayarri, J. O. Berger, Values for Testing Precise Null Hypotheses, The American Statistician 55 (1) (2001) 62–71. doi:10.1198/000313001300339950.

[50] D. Xu, Y. Tian, A Comprehensive Survey of Clustering Algorithms, Annals of Data Science 2 (2) (2015) 165–193. doi:10.1007/s40745-015-0040-1.

[51] H. L. Nguyen, Y. K. Woon, W. K. Ng, A survey on data stream clustering and classification, Knowledge and Information Systems 45 (3) (2015) 535–569. doi:10.1007/s10115-014-0808-1.

[52] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, L. Da Fontoura Costa, A systematic comparison of supervised classifiers, PLoS ONE 9 (4) (2014) e94137. arXiv:1311.0202, doi:10.1371/journal.pone.0094137.

[53] D. M. Abd El-Rehim, G. Ball, S. E. Finder, E. Rakha, C. Paish, J. F. R. Robertson, D. Macmillan, R. W. Blamey, I. O. Ellis, High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses, International Journal of Cancer 116 (3) (2005) 340–350. doi:10.1002/ijc.21004.