# Feature Level Ensemble Method for Classifying Multi-media Data

Saleh Alyahyan and Wenjia Wang

School of Computing Sciences
University of East Anglia, Norwich, UK
{s.alyahyam,w.wang}@uea.ac.uk

**Abstract.** *Multimedia data consists of several different types of data, such as numbers, text, images, audio etc. and they usually need to be fused or integrated before analysis. This study investigates a feature-level aggregation approach to combine multimedia datasets for building heterogeneous ensembles for classification. It firstly aggregates multimedia datasets at feature level to form a normalised big dataset, then uses some parts of it to generate classifiers with different learning algorithms. Finally it applies three rules to select appropriate classifiers based on their accuracy and/or diversity to build heterogeneous ensembles. The method is tested on a multimedia dataset and the results show that the heterogeneous ensembles outperform the individual classifiers as well as homogeneous ensembles. However, it should be noted that, it is possible on some cases that the combined dataset does not produce better results than using single media data.*

**Keywords:** Multimedia data mining, Feature level data aggregation, Diversity, Heterogeneous ensemble, Classification

## 1 Introduction

There has been a rapid rise in the generation of multimedia data, not merely in terms of quantity, but also in the level of complexity as well as variety. Dealing with a vast volume and variety of multimedia data presents a considerable challenge to the existing techniques in machine learning and data mining. Therefore it becomes necessary to develop new techniques and methods that are capable of dealing with multimedia data more effectively and efficiently for various data mining tasks, e.g. classification. Classification of multimedia data has numerous important applications in a wide range of fields, including crime detection, healthcare and business, etc. Two phases are usually needed for classifying multimedia data: first, features need to be extracted from various media datasets and aggregated; second, suitable machine learning algorithms need to be applied[1] to generate classifiers.

In general, multimedia data are characterised by some kinds of heterogeneity, as they often contain different types of data, such as text, images [2], video and audio. These characteristics present an opportunity to apply machine learning methods with two different strategies. The first is to combine the multiple media datasets with their features, which is usually called as feature level data fusion. It involves extracting the

features from different types of data from all the data sources, and combining or merging all the features in order to generate just one dataset, which is sometimes referred as a single flat dataset. The second strategy is called as decision-level data fusion. Instead of fusing all the subsets into one big set in feature-level fusion, this strategy uses each subset separately to generate models, and then combines the output decision of the models to produce a final decision[3]. This study will focus on the former strategy as it is commonly used in data mining practice to investigate how ensemble methods and feature–level fusion could be used more effectively to improve accuracy of classification.

An ensemble combines multiple models using a grating technique with an aim of improving results[4], which has been demonstrated beneficial in the machine learning field for the problems with single media data. However, for multimedia data, there are various factors, including the individual model accuracy, diversity among member models, the number of member models and the decision fusion function used in the ensemble [5] [6] [7] [8] [9], which need to be taken into consideration in order to build an effective ensemble.

This research investigates the methods for building effective ensembles for feature–level fused multimedia data, particularly the heterogeneous ensembles, which are composed of different types, such as decision trees, Bayesian networks and neural networks etc., of classifiers, to examine if an heterogeneous ensemble is more accurate and reliable than homogeneous ensembles – composed of classifiers of the same type, e.g. decision trees only.

The rest of the paper is organized as follows. Section 2 briefly reviews some related previous studies. Section 3 describes our proposed methods in detail, including the tools and programs used in the research. Section 4 provides details of the experiment conducted and our results. Section 5 gives conclusions and suggestions for the further work.

## 2 Related Work

There are several studies that have applied machine learning methods to multimedia data. Aalaa et al [10] applied the machine learning clustering method on heterogeneous (not necessarily multimedia though) datasets. Due to the fact that there were not many heterogeneous datasets publicly available, they created their own heterogeneous datasets, which contained different types of media. Their combined data achieved a significant advantage on clustering performance to that of using only one type of data.

Tuarob et al [11] applied the machine learning heterogeneous ensemble approach to classify social media datasets. They conducted their experiments using three datasets: two datasets collected from Twitter and one from Facebook. They used five different feature extraction methods to generate the data needed for machine learning algorithms. Each of them created a subset of all the combined data. Five base classifiers were used in their experiments, and the classifiers results were combined using different ways, including majority voting and weighted voting. They suggested that the additional features may increase the accuracy of classifiers. However, strictly speaking, in this study, the datasets are not of multimedia, but a single media of multiple textual datasets.

Mehmood and Rasheed [12] classified microbial habitat preferences, based on codon/ bi-codon usage. They attained a high dimensional data set by combining different datasets from different data sources. They showed that the combination, on the feature level, leads to a high dimensional dataset. Thus, they focused on feature selection to reduce the dimensionality of the combined dataset. They reduced a huge number of variables with accepted classification accuracy.

Chen et al [13] also conducted an experiment on combining heterogeneous datasets to a single dataset, and applied homogeneous ensemble classification methods upon it. They used support vector machine as base classifier. In addition, they used real-word microblog datasets, provided by Tencent Weibo. Their results show that the aggregated dataset outperforms any single dataset. Nevertheless, the datasets they used are not of multimedia either and hence how effective their ensemble methods on multimedia data is unknown.

In summary, previous studies have used feature-level combination methods and different machine learning approaches to analyse so-called heterogeneous datasets, whilst in fact their datasets mostly come from different data sources of the same type. Thus, these studies were limited by their single medial of data and how their methods may perform on multimedia datasets is unknown.

## 3 The Feature Level Ensemble Method

### 3.1 The Framework of the Feature Level Ensemble Method

The proposed feature–level ensemble method (FLEM), as illustrated in Figure 1, consists of four modules/stages: multimedia data aggregation module, modelling module, model selection module and combination module.

In general,a multimedia dataset (MMD) should consist of several subsets of various media, e.g. text, images, audio, etc. The FLEM starts with extracting $D_i$'s features ($1 \leq i \leq n$), from each subset of the MMD by using appropriate feature extraction methods. Then, all features are normalised and aggregated to form one big dataset, i.e, $D = N(D_1 \cup D_2 \cup D_3 \cup ... \cup D_n)$. These operations are usually called feature aggregation, which is why our approach is named as Feature–Level Ensemble Method, or FLEM in short.

The second stage is to generate various types of individual models, $m_i$ ($1 \leq i \leq n$), to create a pool of models, $PM = \{m_1, m_2, ..., m_n\}$ as the member candidates of ensemble. The models are called homogeneous models if they are generated by using the same learning algorithm with variations on its parameters and/or data partitions, or called heterogeneous models if they are generated by using different algorithms. A homogeneous ensemble is built with just homogeneous models, whilst a heterogeneous ensemble is constructed with heterogeneous models. In this study, over 10 different base learning algorithms have been selected to generate homogeneous and heterogeneous individual models.

The third stage involves model selection based on a set of defined criteria and rules. In this study, *accuracy* and *diversity* are used as selection criteria either separately or

jointly. Three different rules for model selection are devised explained in the next section. Finally, the selected models are combined into one ensemble and their classification decisions are aggregated using a combination method to reach the final form of the ensemble.

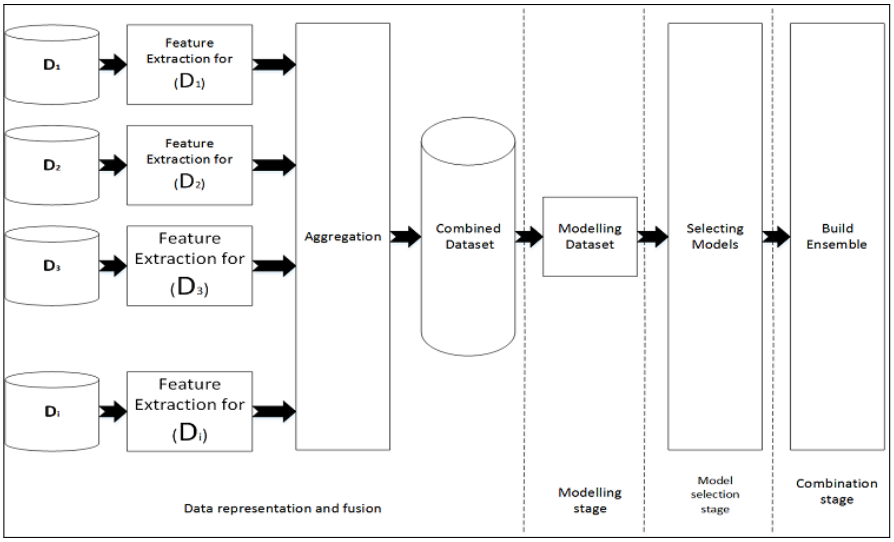


**Fig. 1.** A general framework for the feature–level ensemble method(FLEM).

### 3.2 Rules for selecting models

Three rules, $R0, R1$ and $R2$, as illustrated by Fig. 2, were devised for selecting models based on various criteria.

**R0**: This rule uses *accuracy* of classifiers as the criterion for selection. The FLEM firstly computes the accuracy, $(Acc(m_i))$, for each of the $n$ models in the model pool $PM$ and sort them in a descending order based on the accuracy of the models $(Acc(m_i))$. Then the FLEM selects the $N$ most accurate models from the PM using equation 1 and add them to the ensemble, $\Phi$, as shown in Figure 2(a).

$$m_i = max\left\{Acc(m_j), m_j \in PM\right\} i = 1...N. \quad (1)$$

**R1**: This rule uses both *accuracy* and *diversity* in sequence for model selection. FLEM first removes the most accurate model (MAM), $m_1 = max\left\{Acc(m_j), m_j \in PM\right\}$, from PM and adds it to ensemble $\Phi$.

Then the pairwise diversity between MAM and remaining models in PM, are calculated with the Double Fault (DF) measure[15]. Then FLEM sorts the models in PM in a decreasing order based on the magnitude of the DF's. The $(N-1)$ most diverse models

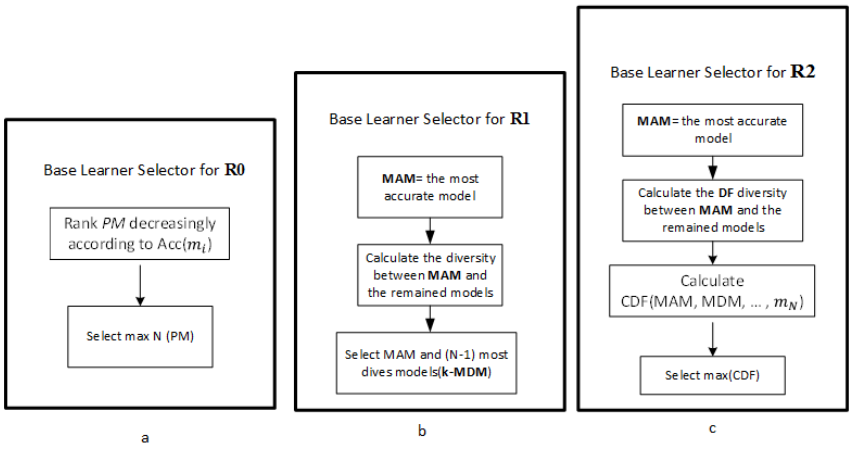


**Fig. 2.** Main steps for R0, R1 and R2 in HES [14]

from the sorted PM are selected (Equation 2) and added to the ensemble, Φ. Therefore Φ now contains MAM and the (N-1) most diverse models from PM.

$$m_i = max\left\{DF(m_1, m_j), m_j \in PM\right\} i = 2...N \tag{2}$$

**R2:** This rule uses both *accuracy* and two types of *diversity* measures, namely the *DF* diversity and the *Coincident Failure Diversity*(CFD) method[16]. Firstly, MAM is selected and removed from PM and added to Φ. Then the most diverse model (MDM) is determined from PM using $MDM = max\left\{DF(m_1, m_j), m_j \in PM\right\}$ and added to Φ, which now contains both the MAM and MDM models. From this point on, the $CFD$ diversity measure is used to select further models with an aim of maximising the CFD diversity of the ensemble if they are included in Φ. The ensemble with the maximum CFD diversity is selected as the final ensemble, Φ using $\Phi = max\left\{CFD(\Phi \Leftarrow m_j), m_j \in PM\right\}$.

Rule *R*2 may be time consuming if the size of the model pool PN is large and all the possible combinations between Φ and the remaining members of PM are considered.

### 3.3 Implementation of FLEM

The FLEM is implemented with Java, based on Weka API. The experiment was carried out on a normal PC with an I7 processor and 16 GB RAM. As FLEM is flexible for selecting candidate classifiers, we selected 10 efferent base classifiers provided in the WEKA library, which are: three types of decision trees (*J48, RandomTree, REP-Tree*), two Bayesian methods (*NaiveBayes, BayesNet*), Support vector machine(SMO), two rule induction methods(*JRip, PART*) and two Lazy learners (*IBk and LWL*).

# 4 Experiment Design and Results

## 4.1 Dataset

We conducted our experiment using a benchmark dataset – 8 Scene Categories Dataset [17], which contains two parts in different media: 2688 images and their annotations represented by XML files. The images are categorized into eight classes in according to their scenes and objects captured by the images. Each XML file contained a number of tags that describe an image. The annotations were dealt with as text and 782 textual features were extracted out from the texts to form a data subset $D_t$. For the imagery data, 567 features were extracted out from the images using Histograms of Oriented Gradients (HOG) [18] to form another data subset, i.e. imagery data $D_g$.

The textual and imagery features from these two data subsets were aggregated to form a single dataset, $D = N(D_t \cup D_g)$, which contains 2688 instances and each has 1358 features in total as inputs and one classification output of 8 classes.

## 4.2 Experiment Design and Results

We conducted a series of experiments to investigate the performance of the FLEM working with three selection rules separately on the multimedia data. The factors that were investigated include (1) the performance measures and criteria for selecting classifiers, which are represented by the three rules: R0, R1 and R2, (2) the size of ensemble – varied from 3, 5, 7 to 9, and (3) the salience of multimedia data, i.e. if the combined multimedia data $D$ can produce better results, compared with each of single-media data subsets: $D_t$ and $D_g$. For each specific set-up, the experiment is repeated 5 times with different data partitions to check consistency.

In addition, for comparison with heterogeneous ensembles, homogeneous ensembles were built with the classifiers selected only from the same type. As ten different types of base learning algorithms were used for generating classifiers, ten homogeneous ensembles were constructed for each set-up of above listed factors.

Therefore, for all possible combinations of these parameters, over 200 sets of experiments were conducted in total.

**Results of FLEMs built with three rules and variable sizes:** Fig 3, 4 and 5 show some results (means and standard deviations) obtained from these experiments of FLEMs constructed with three rules and different sizes.

The results of varying ensemble size from 3 to 9 on the test data for each of the three rules are summarised and shown by Fig. 6. As can be seen, three rules produced quite different ensembles at almost every size.

Rule R0 worked very well with its first ensemble when $N = 3$, after that its accuracy went down continuously, which is not surprising given that it always chooses the best remaining classifiers in the model pool and the classifiers selected after $N = 3$ will be worse and worse than those previously selected. As for Rule R1, in general, it is the worst all way down when using accuracy and diversity measured by the Double–fault rate as two selection criteria in a sequential manner. In contrast, Rule R2 performed the best. It not just started as good as that of R0 when $N = 3$, but also got even better when

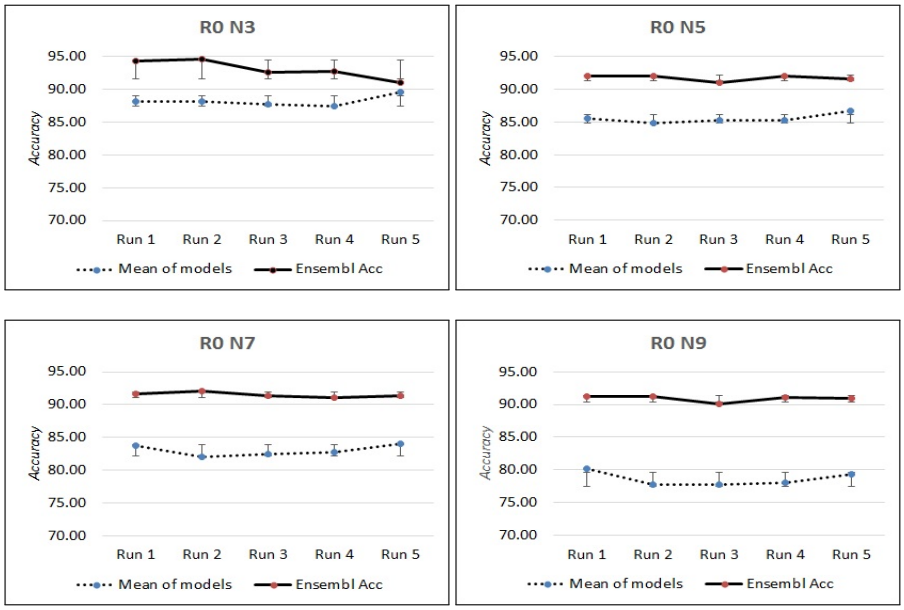


**Fig. 3.** The results of FLEMs built with rule R0. Each sub-graph shows the ensembles with different sizes 3, or 5, or 7 or 9. The solid and dashed lines are the mean accuracy of FLEMs and the mean accuracy of the models in the FLEMs respectively, with their standard deviations as error bars.

the size of ensemble increased to 5 and maintained the best accuracy. It demonstrates that using an appropriate diversity – CFD in this case, when selecting the classifiers, can enhance the performance of heterogeneous ensembles as heterogeneous classifiers are more diverse and hence help to improve the accuracy of these ensembles.

However, This shows the effectiveness of the CFD which we apply model selection for heterogeneous classification ensemble methods.

**Results of using text, images and combined datasets:** As designed, further experiments were conducted by separately using three sets of data: text, imagery and combined, in order to investigate if the aggregation of subsets of multimedia data gives better results. The experiments on the textual dataset $D_t$ alone were conducted with our Heterogeneous Ensemble System, called HES_T, and their results have been reported in our earlier paper[14]. The experiments on the image dataset $D_g$, called HES_G, were conducted in this study in the same way as the one used for the text experiments. The results of HES_G, obtained in these experiments, are shown in Fig. 7, 8and 9.

The summary of the results of varying the ensemble size from 3 to 9 on the test dataset for each of the three rules is shown by Fig. 10.

A further observation from these results is that, using FLEM, the accuracy of the combined text and imagery datasets was lower than that of using the text dataset alone. Furthermore, the accuracy of the image dataset alone was lower than that of the text

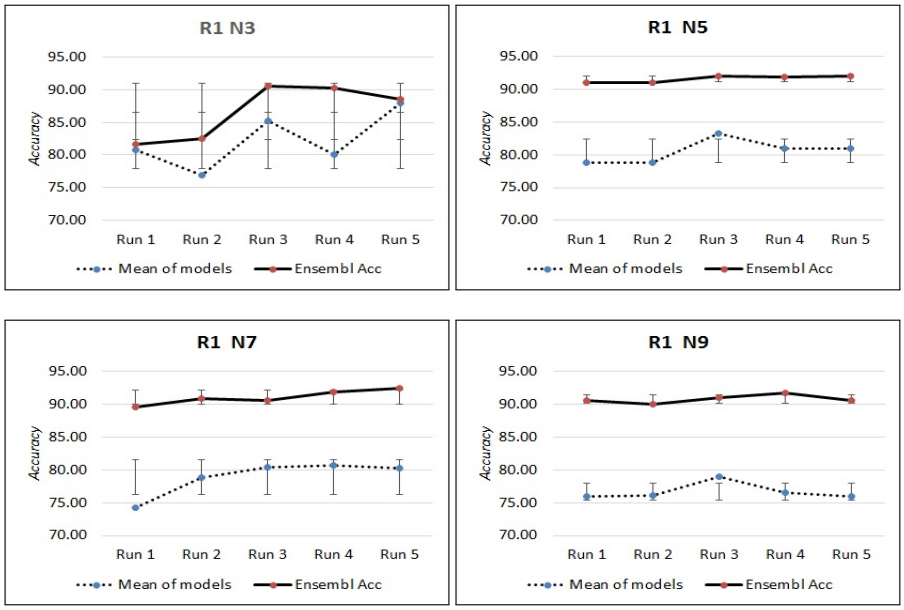


**Fig. 4.** The mean test accuracy of FLEMs built with rule R1. Four sub-graphs show the ensembles with different size of 3, 5, 7 and 9.

dataset, as shown in Fig 11. A plausible explanation for these differences is that the features extracted from the imagery dataset did not very well represent the information associated with the underlying classification knowledge of the problem, or even worse brought in some noises, and hence confused the learning algorithms to produce quite weak or bad models, which in turn resulted in weak ensembles. On the other hand, the text data, or more precisely speaking the features extracted from the text data, are more representative or salient as the ensembles built with the models trained with the text data are more accurate, about 15% higher than those of ensembles built with the image data.

With the combined dataset, the ensembles produced variable accuracies, depending on the rules used. The results of the FLEMs built with R0 are the best and comparable to those built with the text data only. But the whole, in this application, the aggregation of multimedia datasets did not offer much additional benefit in terms of improving classification accuracy.

**Comparison with homogeneous ensembles:** Another set of experiments was conducted to compare the performance between heterogeneous and various homogeneous ensembles, built using all the three model selection rules that have been implemented in FLEMs.

Table 1 shows the mean accuracies for the feature level heterogeneous ensembles and all homogeneous ensembles built with rules R0, R1 and R2. It is very clear that,

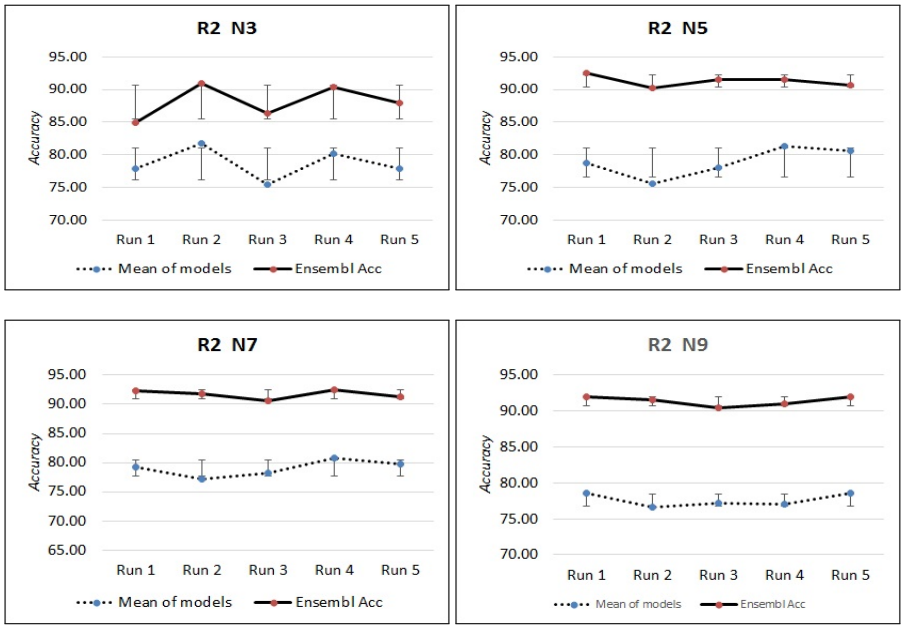


**Fig. 5.** The mean test accuracy of FLEMS built with rule R2. Four sub-graphs show the ensembles with different size of 3, 5, 7 and 9.

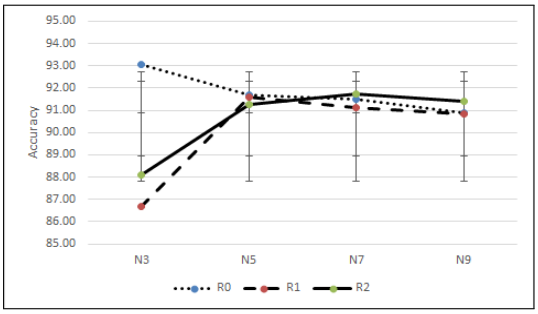


**Fig. 6.** Comparison of three rules as the size of FLEM varies.

on average the heterogeneous ensembles are much better, about 10% higher, than the homogeneous ensembles constructed using the same rules.

## 5 Conclusion and Future Work

Aggregating and mining multi-media datasets effectively is a challenge task in machine learning and data mining fields. In this work, we developed a feature–level ensemble method (FLEM) with an aim of achieving better classification of multimedia data. Our

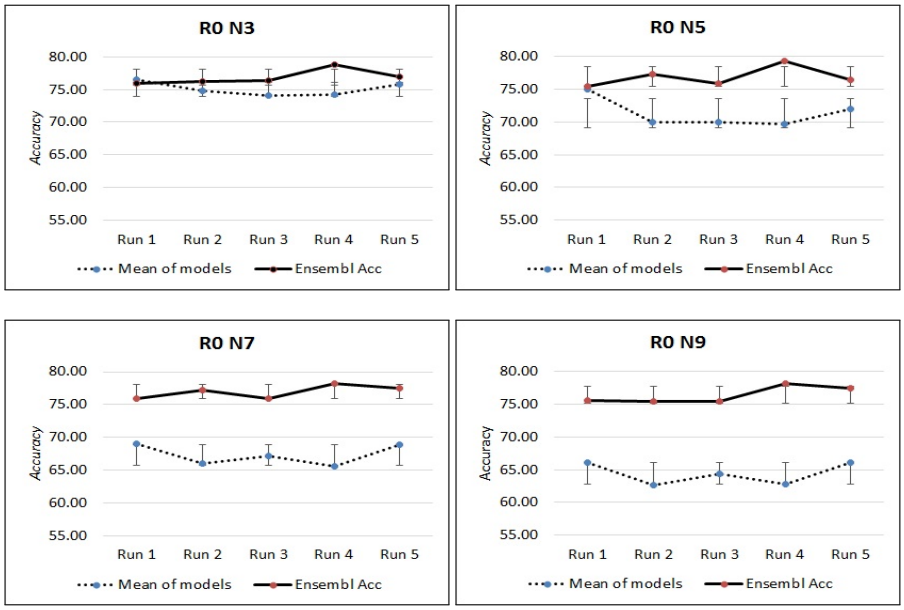


**Fig. 7.** The results of HESs built with rule R0 and different sizes (3, 5, 7 and 9) for the image dataset. The solid and dashed lines are the mean accuracy of HESs and the models of the HESs, with standard deviations as error bars, respectively.

**Table 1.** The comparison results between the heterogeneous ensemble and the homogeneous ensemble for FLEM for all the three rules

| | Heterogeneous | | Homogeneous | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| R0 | 91.79 | 0.92 | 80.05 | 0.78 |
| R1 | 90.05 | 2.26 | 80.04 | 0.99 |
| R2 | 90.63 | 1.70 | 80.01 | 0.97 |

FLEM consists of four stages: extracting features from multimedia subsets and aggregating them into a single dataset, modelling the combined dataset, selecting models with different rules based on various criteria, and building heterogeneous ensembles. The experimental results have demonstrated it is capable of handling multimedia datasets – unstructured text data and imagery data, simultaneously and builds the best ensembles with appropriate datasets, with either combined multi-media data or single-media data. In general, the heterogeneous ensembles are much better than homogeneous ensembles in terms of accuracy and consistency.

Another point drawn from the results of this study is that it should be cautious when combining multiple data subsets of a problem because the aggregated data may not produce a better result than that of using data subsets of single–media. Possible reasons include poor features extracted from each subset, which capture more noise rather than useful information; and/or inappropriate aggregation, which may introduce

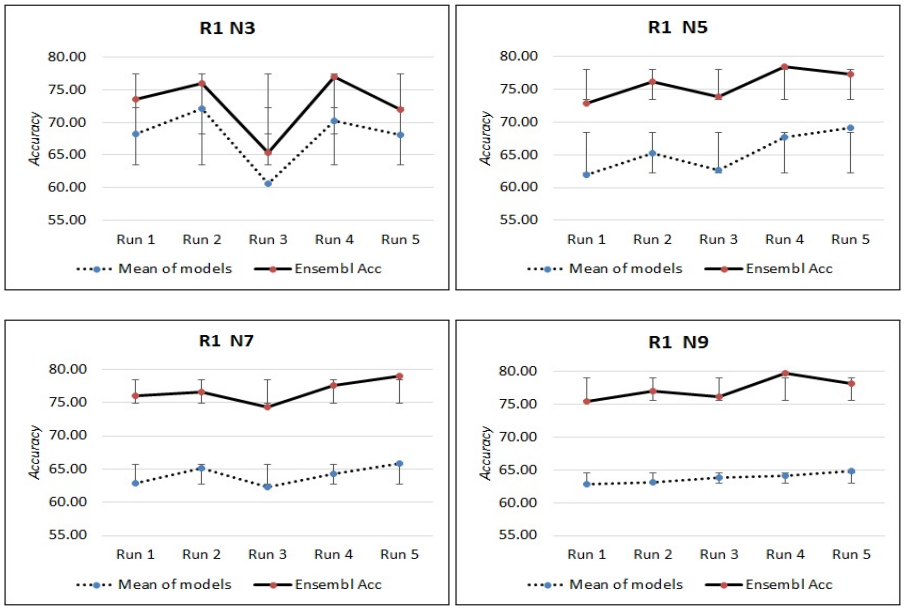


**Fig. 8.** The results of HESs built with rule R1 and different sizes (3, 5, 7 and 9) for the image dataset.

some inconsistency or even contradictions into the final dataset and therefore cause a great deal of difficulty and/or confusion in learning.

Some further work can be done in our approach on various aspects. Fro example, firstly, it could be useful to apply some feature selection methods on each of data subsets before aggregation, to eliminate irrelevant or redundant features, which in turn can reduce the dimensionality of the data and simplify learning. Secondly, more rules could be devised to select the models to increase overall accuracy levels. Thirdly, it would prove useful to analyse multi-media datasets which contain other, different types of media, which have not yet been the subject of this research. Finally, instead of aggregating multiple data sets at feature–level, it appears more promising to apply decision–level aggregation strategy, which is to generate the models independently from each of data subsets and then combine them to form an ensemble.

## References

1. Z. Zhang and R. Zhang, "Multimedia data mining," *Data Mining and Knowledge Discovery Handbook*, pp. 1081–1109, 2010.
2. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *ieee transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.
3. C. Ballard and W. Wang, "Dynamic ensemble selection methods for heterogeneous data mining," in *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*, pp. 1021–1026, IEEE, 2016.

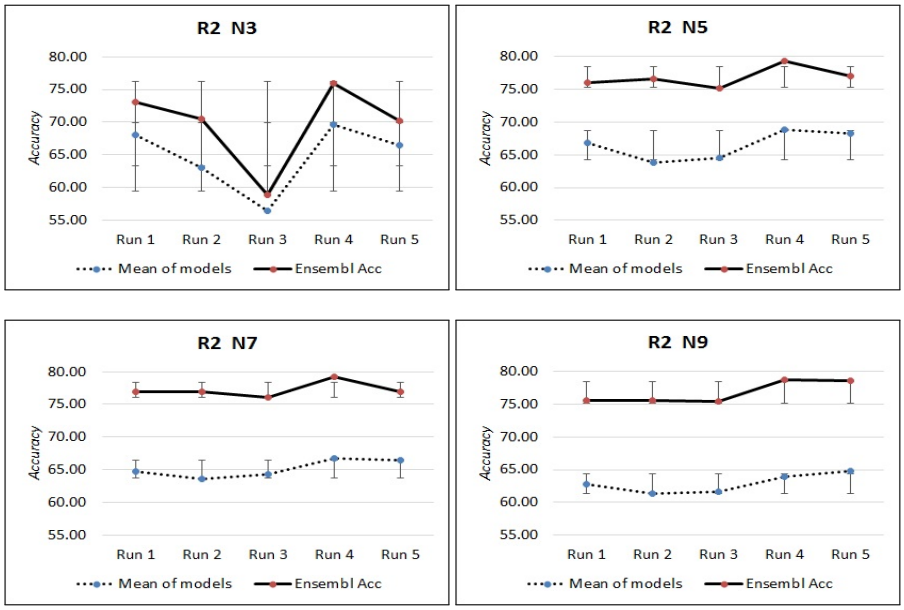


**Fig. 9.** The results of HESs built with rule R2 and different sizes (3, 5, 7 and 9) for the image dataset.

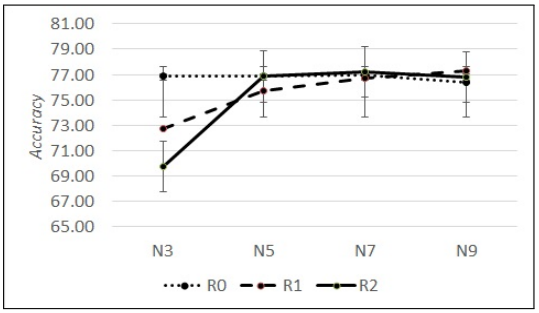


**Fig. 10.** Comparing all three rules in four different sizes of the HESs for the image dataset only.

4. T. G. Dietterich, *Ensemble methods in machine learning*, pp. 1–15. Springer, 2000.
5. W. Wang, "Some fundamental issues in ensemble methods," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pp. 2243–2250, IEEE, 2008.
6. R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the twenty-first international conference on Machine learning*, p. 18, ACM, 2004.
7. S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox, "Ensembles of models for automated diagnosis of system performance problems," in *Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on*, pp. 644–653, IEEE, 2005.

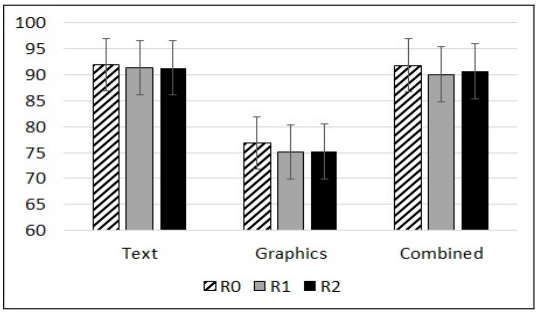


**Fig. 11.** Comparison of all the ensembles built with three rules for text dataset, image dataset and the combined multimedia dataset respectively.


8. G. Zenobi and P. Cunningham, *Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error*, pp. 576–587. Springer, 2001.
9. Y. Liu, X. Yao, and T. Higuchi, "Evolutionary ensembles with negative correlation learning," *Evolutionary Computation, IEEE Transactions on*, vol. 4, no. 4, pp. 380–387, 2000.
10. A. Mojahed, J. H. Bettencourt-Silva, W. Wang, and B. de la Iglesia, "Applying clustering analysis to heterogeneous data using similarity matrix fusion (smf)," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 251–265, Springer, 2015.
11. S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram, "An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages," *Journal of biomedical informatics*, vol. 49, pp. 255–268, 2014.
12. T. Mehmood and Z. Rasheed, "Multivariate procedure for variable selection and classification of high dimensional heterogeneous data," *Communications for Statistical Applications and Methods*, vol. 22, no. 6, pp. 575–587, 2015.
13. Z.-Y. Chen, Z.-P. Fan, and M. Sun, "Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data," *European Journal of Operational Research*, vol. 241, no. 2, pp. 422–434, 2015.
14. S. Alyahyan, M. Farrash, and W. Wang, "Heterogeneous ensemble for imaginary scene classification," in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR, Porto - Portugal, November 9 - 11, 2016.*, pp. 197–204, 2016.
15. G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9, pp. 699–707, 2001.
16. D. Partridge and W. Krzanowski, "Software diversity: practical statistics for its measurement and exploitation," *Information and software technology*, vol. 39, no. 10, pp. 707–717, 1997.
17. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.
18. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.