MIRACLE at ImageCLEFanot 2007: Machine Learning Experiments on Medical Image Annotation

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3]
José Carlos González-Cristóbal[1,3], José Miguel Goñi-Menoyo[1]
[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid.
[3] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, jvillena@daedalus.es
josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

Abstract

This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Annotation task of ImageCLEF 2007. Our areas of expertise do not include image analysis, thus we approach this task as a machine-learning problem, regardless of the domain.

FIRE is used as a black-box algorithm to extract different groups of image features that are later used for training different classifiers in order to predict the IRMA code. Three types of classifiers are built. The first type is a single classifier that predicts the complete IRMA code. The second type is a two level classifier composed of four classifiers that individually predict each axis of the IRMA code. The third type is similar to the second one but predicts a combined pair of axes. The main idea behind the definition of our experiments is to evaluate whether an axis-by-axis prediction is better than a prediction by pairs of axes or the complete code, or vice versa.

We submitted 30 experiments to be evaluated and results are disappointing compared to other groups. However, the main conclusion that can be drawn from the experiments is that, irrespective of the selected image features, the axis-by-axis prediction achieves more accurate results not only than the prediction of a combined pair of axes but also, in turn, than the prediction of the complete IRMA code. In addition, data normalization seems to improve the predictions and vector-based features are preferred over histogram-based ones.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital libraries.

Keywords

Information Retrieval, medical image, image annotation, classification, IRMA code, axis, learning algorithms, nearest-neighbour, machine learning.

1. Introduction

The MIRACLE team is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise (SME) founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks, including the main bilingual, monolingual and cross lingual tasks as well as in ImageCLEF [7], Question Answering, WebCLEF and GeoCLEF tracks.

This paper describes our second participation in the ImageCLEF Medical Image Annotation task of ImageCLEF 2007. Briefly, the objective of this task (fully described in [6]) is to provide the IRMA (Image Retrieval in Medical Applications) code [5] for each image of a given set of 1,000 previously unseen medical (radiological) images covering different medical pathologies. 10,000 classified training images are provided to be used in any way to train a classifier. This task uses no textual information, but only image-content information. We approach this task as a machine learning problem, regardless of the domain, as our areas of expertise do not include image analysis research [4].

2. Description of Experiments

FIRE (Flexible Image Retrieval Engine) [2] [3] is a freely available content-based information retrieval system developed under the GNU General Public License that allows to perform query by example on images, using an image as the starting point for the search process and relying entirely on the image contents. FIRE offers a wide repertory of available features and distance functions. Specifically, the distribution package includes a set of scripts that extracts different types of features from the images, including color/gray histograms, invariant features histograms, Gabor features, global texture descriptor, Tamura features, etc.

Our approach to the task is to build different classifiers that use image features to predict the IRMA code. For that purpose, all images in the training, development and testing dataset have been processed with FIRE. The extracted features have been arranged in three groups, as shown in **Table 1**, to build the training data matrixes for the classifiers.

Table 1. Training data matrixes.

Name[(1)]	FIRE – Image Features	Dimension[(2)]
Histogram	Gray histogram and Tamura features	768
Vector	Aspect ratio, global texture descriptor and Gabor features	75
Complete	Gray histogram, Tamura features, aspect ratio, global texture descriptor and Gabor features	843

(1) Used in the experiment description

(2) Number of columns of the matrix; the number of rows is 10,000 for the training dataset and 1,000 for the development and testing dataset.

Different strategies have been evaluated, using several multiclassifiers built up with a set of specialized individual classifiers:

- **IRMA Code Classifier**: single classifier that uses the image features to predict the complete IRMA code (4 axes: Technical, Direction, Anatomical and Biological).
- **IRMA Code Axis Classifier**: a two level classifier that is composed of four different classifiers that individually predict the value of each axis of the IRMA code; the prediction is the concatenation of partial solutions.
- **IRMA Code Combined Axis Classifier**: similar to the axis classifier, this one predicts the axes grouped in pairs.

These classifiers are all based on the K-Nearest-Neighbour algorithm [8], with K=10, to predict the output class.

The main idea behind the definition of the experiments is to evaluate whether an axis-by-axis prediction is better than a prediction by pairs of axes or the complete code, or vice versa. In addition, the effect of applying the data normalization will be also analyzed.

Finally we submitted 30 experiments to be evaluated, described in **Table 2**.

Table 2. Experiment set.

Run Identifier	Features	Prediction[(1)]	Normalization[(2)]
MiracleA	Complete	Complete code	NO
MiracleAA	Complete	Axis-by-axis	NO
MiracleAATABD	Complete	Combined axis: T+A and B+D	NO
MiracleAATBDA	Complete	Combined axis: T+B and D+A	NO
MiracleAATDAB	Complete	Combined axis: T+D and A+B	NO
MiracleH	Histogram	Complete code	NO
MiracleHA	Histogram	Axis-by-axis	NO
MiracleHATABD	Histogram	Combined axis: T+A and B+D	NO
MiracleHATBDA	Histogram	Combined axis: T+B and D+A	NO
MiracleHATDAB	Histogram	Combined axis: T+D and A+B	NO
MiracleV	Vector	Complete code	NO
MiracleVA	Vector	Axis-by-axis	NO

MiracleVATABD	Vector	Combined axis: T+A and B+D	NO
MiracleVATBDA	Vector	Combined axis: T+B and D+A	NO
MiracleVATDAB	Vector	Combined axis: T+D and A+B	NO
MiracleAn	Complete	Complete code	YES
MiracleAAn	Complete	Axis-by-axis	YES
MiracleAATABDn	Complete	Combined axis: T+A and B+D	YES
MiracleAATBDAn	Complete	Combined axis: T+B and D+A	YES
MiracleAATDABn	Complete	Combined axis: T+D and A+B	YES
MiracleHn	Histogram	Complete code	YES
MiracleHAn	Histogram	Axis-by-axis	YES
MiracleHATABDn	Histogram	Combined axis: T+A and B+D	YES
MiracleHATBDAn	Histogram	Combined axis: T+B and D+A	YES
MiracleHATDABn	Histogram	Combined axis: T+D and A+B	YES
MiracleVn	Vector	Complete code	YES
MiracleVAn	Vector	Axis-by-axis	YES
MiracleVATABDn	Vector	Combined axis: T+A and B+D	YES
MiracleVATDABn	Vector	Combined axis: T+B and D+A	YES
MiracleA	Vector	Combined axis: T+D and A+B	YES

(1) IRMA code axes are: Technical (**T**), Direction (**D**), Anatomical (**A**) and Biological (**B**).

(2) Normalized to range [0, 1].

3. Results

Results are shown in **Table 3**. The "Error count" column contains the experiment score as computed by the task organizers [1]. This score is defined to penalize wrong decisions that are easy to take (i.e., there are few possible choices at that node) over wrong decisions difficult to take (i.e., there are many possible choices at that node). Furthermore, it also penalizes wrong decisions at an early stage in the code (higher up in the IRMA code hierarchy) over wrong decisions at a later stage (lower down in the hierarchy). The "Well-Classified" column shows the actual number of images with correct predicted codes.

Table 3. Evaluation of experiments

Run Identifier	Error count	Well-Classified
MiracleAAn	***158.82**	497
MiracleVAn	159.45	504
MiracleAATDABn	160.25	501
MiracleAATABDn	162.18	499
MiracleVATDABn	174.99	***507**
MiracleAATBDAn	177.60	487
MiracleAATDAB	186.99	450
MiracleHATDAB	187.42	450
MiracleAA	188.93	445
MiracleAATABD	189.21	450
MiracleHA	189.37	445
MiracleHATABD	189.45	450
MiracleHATDABn	189.60	427
MiracleHAn	190.59	428
MiracleHATABDn	195.27	425
MiracleHATBDA	197.10	454
MiracleAATBDA	197.12	453
MiracleH	198.15	459
MiracleA	198.67	458
MiracleHATBDAn	199.40	434
MiracleVATBDAn	221.34	257
MiracleAn	245.92	438

MiracleVATABDn	245.95	234
MiracleVATBDA	303.00	173
MiracleHn	323.66	328
MiracleVA	325.89	148
MiracleVATABD	350.21	110
MiracleVATDAB	419.66	156
MiracleVn	490.66	174
MiracleV	505.62	132

According to the weighted error count score, the best experiment is the one with data normalization that predicts each axis individually using all image features ("histogram" and "vector"). However, considering the number of correctly classified images, the best experiment is the one that uses normalized vector-based features and predicts the combined axis Technical+Direction and Anatomical+Biological.

Figure 1 allows to compare the predictions of the complete IRMA code versus the axis-by-axis predictions. Other similar comparisons are also included in the appendix. The main conclusion to be drawn is that, regardless of the selected image features, the axis-by-axis prediction achieves more accurate results not only than the prediction of a combined pair of axes but also than the prediction of the complete code.

In addition, data normalization seems to improve the predictions and vector-based features are preferred over histogram-based ones.

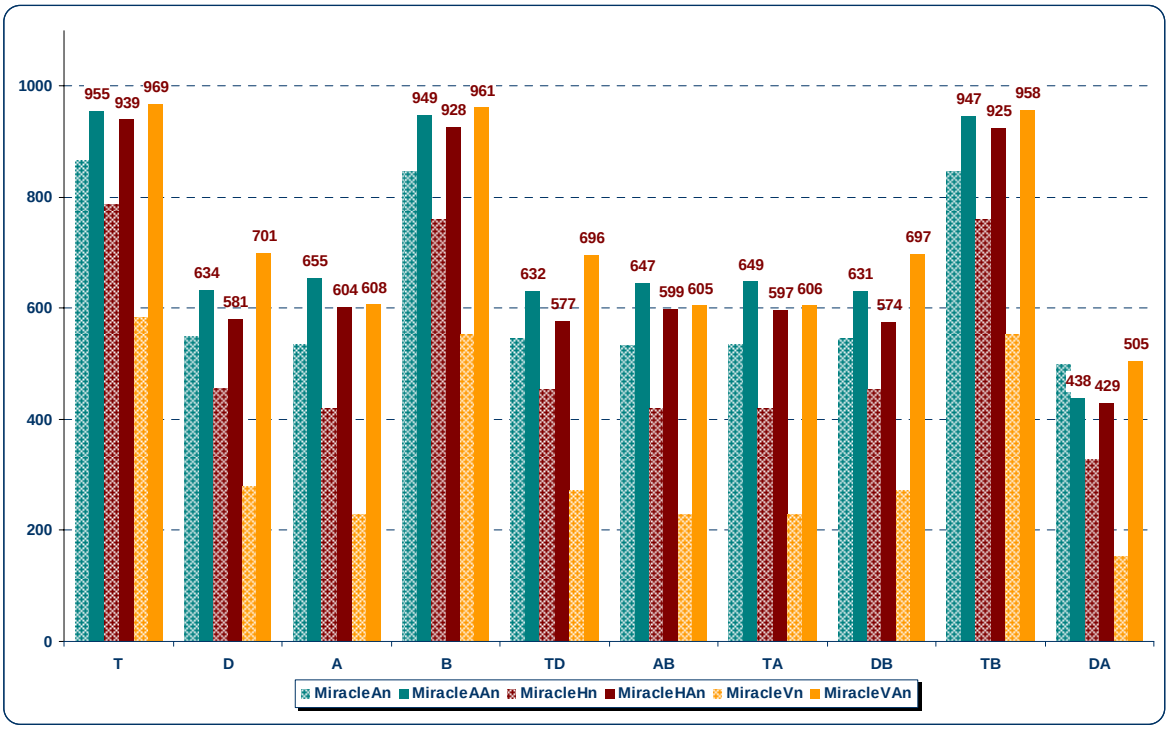

Figure 1. Complete code prediction vs axis-by-axis prediction.

Comparing to other groups, our results were considerably worse. The best experiment reached a score of 26.84, 17% of our own best error count. MIRACLE ranked 9th out of 10 participants in the task.

4. Conclusions and Future Work

The main conclusion that can be drawn from the evaluation is that, irrespective of the selected image features, the best experiments are those that predict the IRMA code from the individual partial predictions of the 1-axis classifiers. Moreover, the predictions of combined pairs of axes are better than the predictions of the complete IRMA code. By extension, it could be concluded that the finer granularity of the classifier, the more accurate

predictions are achieved. In the extreme case, the prediction may be built up from 13 classifiers, one per each character of the IRMA code. This issue will be further investigated and some experiments are already planned.

One of the toughest challenges to face when designing a classifier is the selection of the vector of features that best captures the different aspects that allow to distinguish one class from the others. Obviously, this requires an expert knowledge of the problem to be solved, which we currently lack. We are convinced that one of the weaknesses of our system is the feature selection. Therefore more effort will be invested in improving this topic for future participations.

Acknowledgements

This work has been partially supported by the Spanish R&D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by the Madrid's R&D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

[1] Deselaers, Thomas; Kalpathy-Cramer, Jayashree; Müller, Henning; Deserno, Thomas. Hierarchical classification for ImageCLEF 2007 Medical Image Annotation. On line http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/hierarchical.pdf [Visited 10/08/2007].

[2] Deselaers, T.; Keysers; D.; Ney, H. FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In CLEF 2004, LNCS 3491, Bath, UK, pp 688-698, September 2004.

[3] FIRE: Flexible Image Retrieval System. On line http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html [Visited 10/08/2007].

[4] Goodrum, A.A. Image Information Retrieval: An Overview of Current Research. Informing Science, Vol 3(2), pp 63-66, 2000.

[5] IRMA project: Image Retrieval in Medical Applications. On line http://www.irma-project.org/ [Visited 10/08/2007].

[6] Müller, Henning; Deselaers, Thomas; Kim, Eugene; Kalpathy-Cramer, Jayashree; Deserno, Thomas; Clough, Paul; Hersh, William. Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.

[7] Villena-Román, J.; González-Cristóbal, J.C.; Goñi-Menoyo, J.M.; and Martínez Fernández, J.L. MIRACLE's Naive Approach to Medical Images Annotation. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.

[8] Witten, Ian H.; Frank, Eibe. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Appendix

The following figures compare the predictions of the complete IRMA code versus partial predictions of combined pairs of axes. Only normalized datasets are shown because they lead to better results.

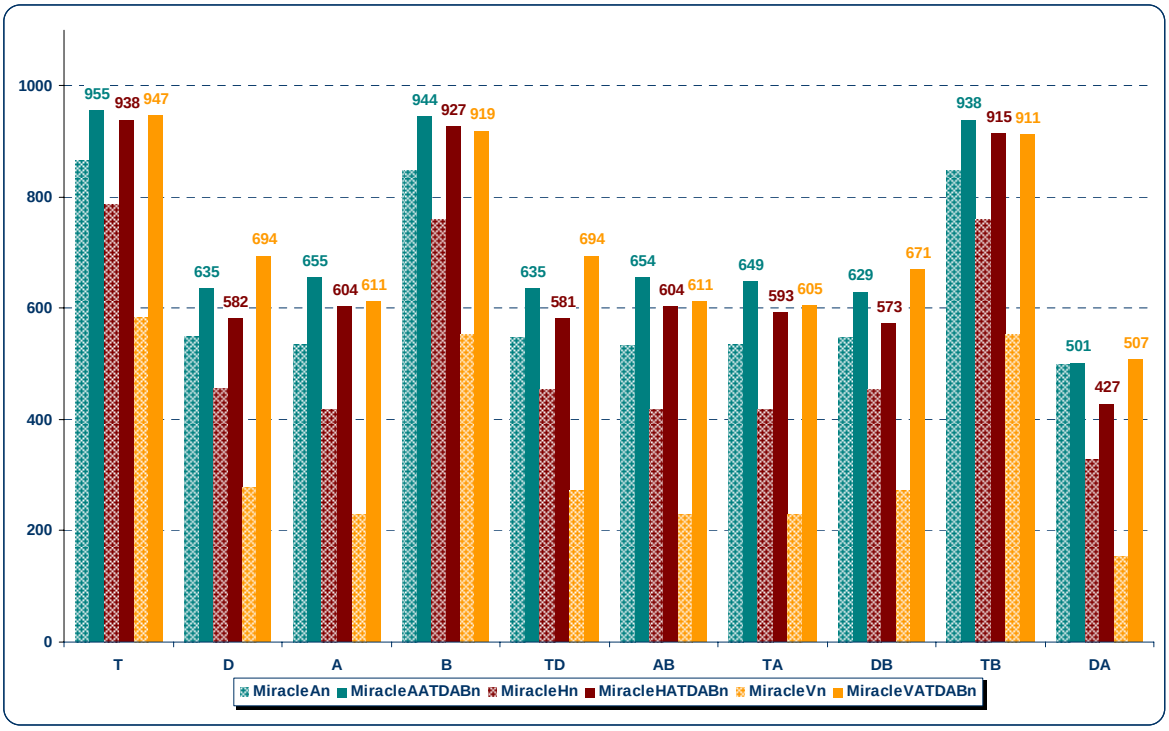

Figure 2. Complete code prediction vs TD+AB combined axis prediction.

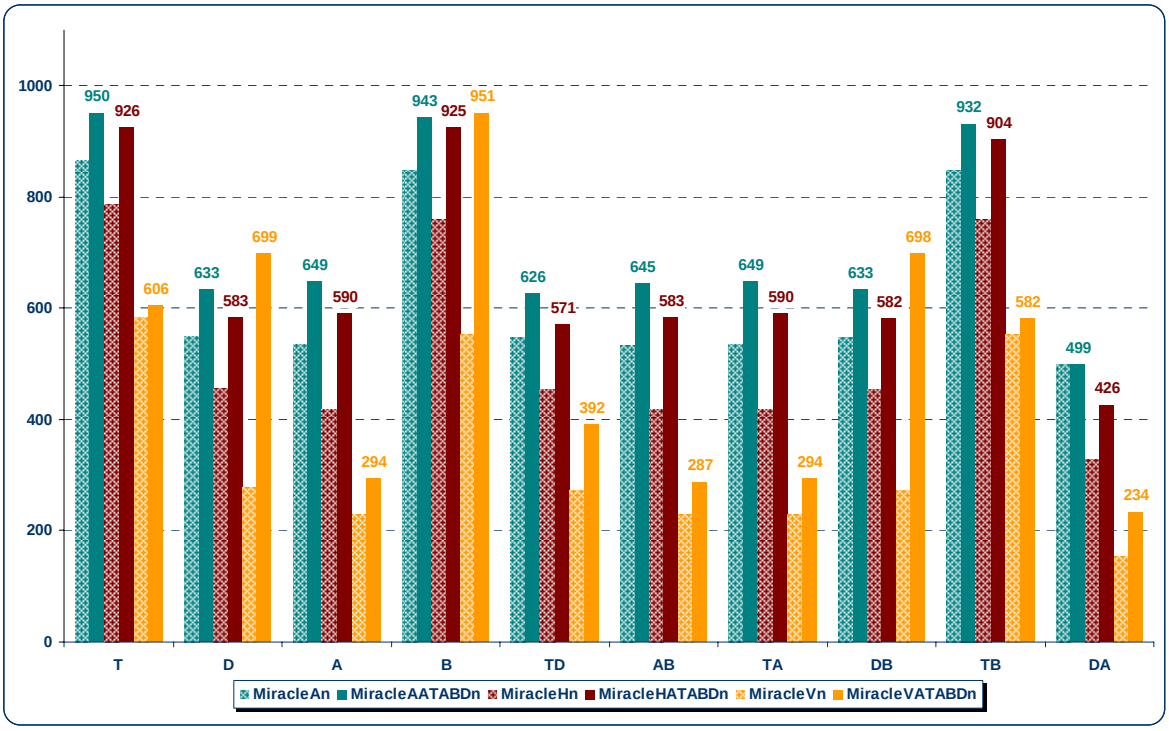

Figure 3. Complete code prediction vs TA+BD combined axis prediction.

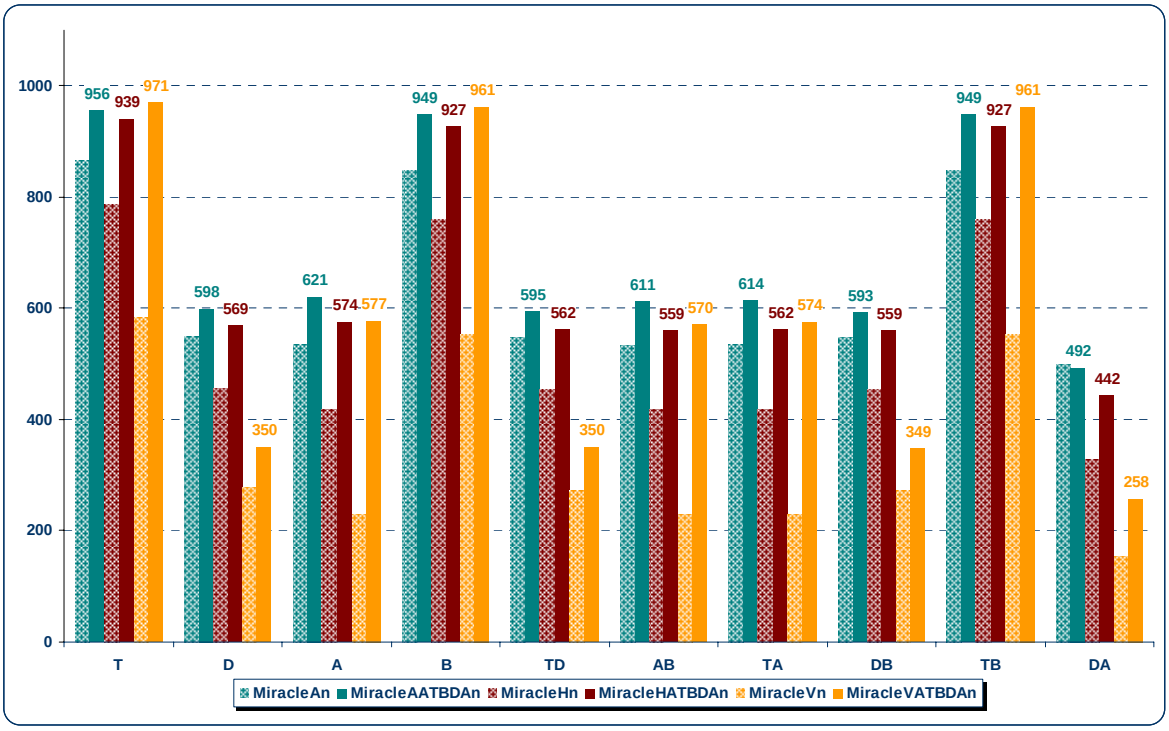

Figure 4. Complete code prediction vs TB+DA combined axis prediction.

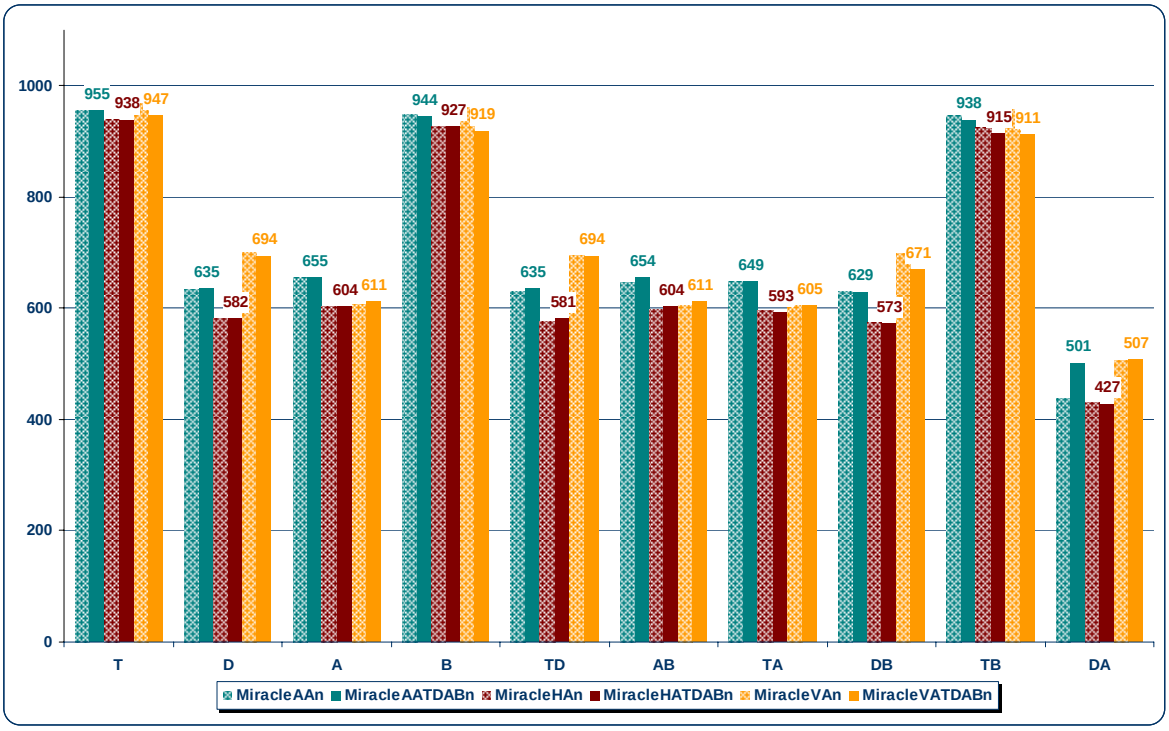

Figure 5. Axis-by-axis prediction vs TD+AB combined axis prediction.

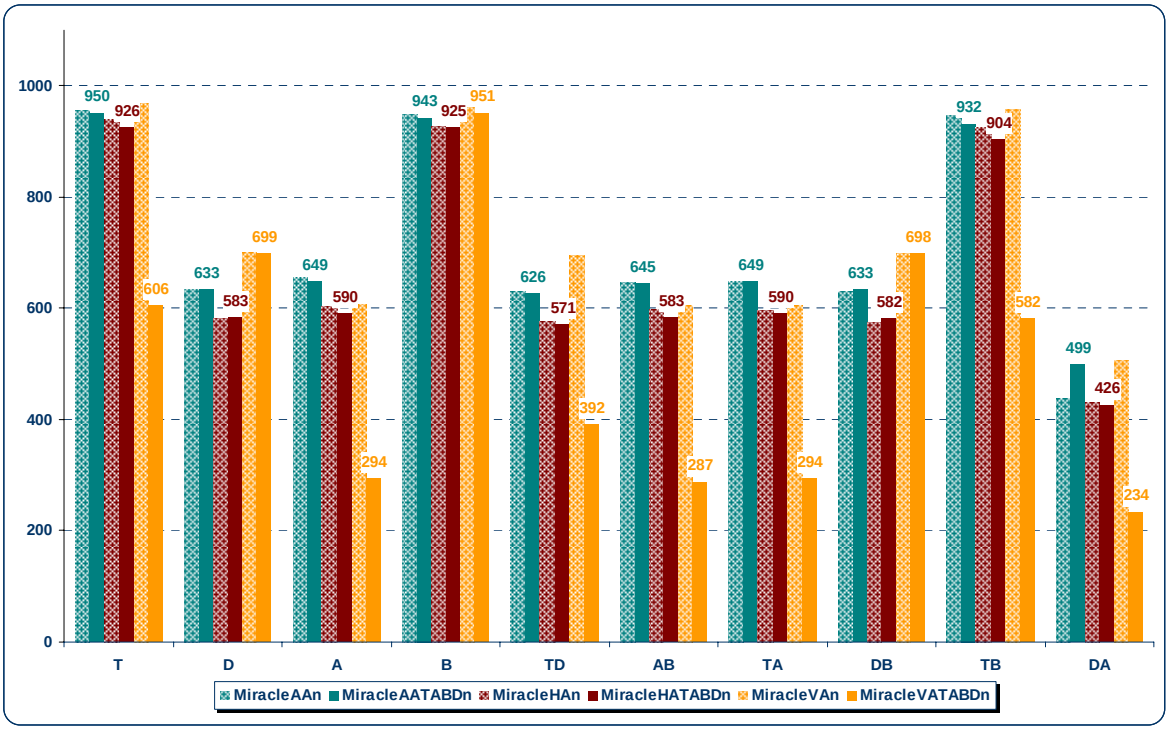

Figure 6. Axis-by-axis prediction vs TA+BD combined axis prediction.

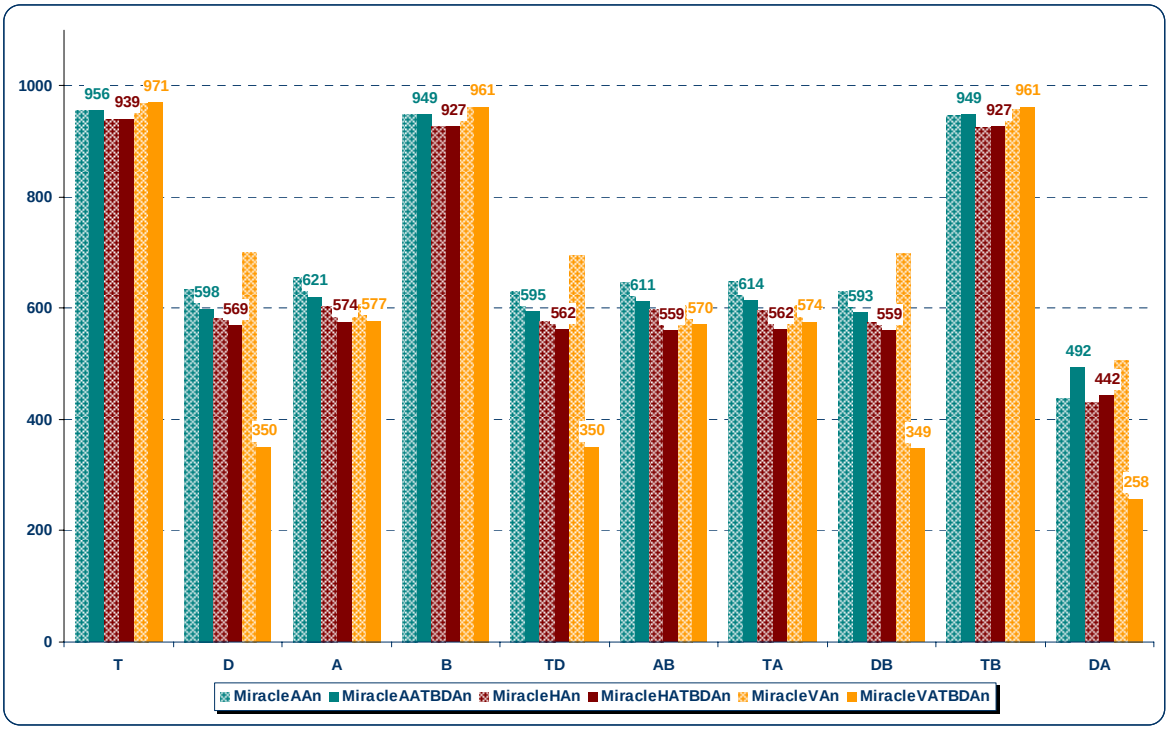

Figure 7. Axis-by-axis prediction vs TB+DA combined axis prediction.