Caesarea Rothschild Institute
for Interdisciplinary Applications
of Computer Science

UNIVERSITY OF HAIFA

Using Machine learning to Identify Patients with Parkinson Disease

Hananel Hazan
Department of Computer Science
University of Haifa, Israel
hhazan01@cs.haifa.ac.il

Dan Hilu
Department of Computer Science
University of Haifa, Israel
danram4@zahav.net.il

Larry M. Manevitz
Department of Computer Science
University of Haifa, Israel
manevitz@cs.haifa.ac.il

Shimon Sapir
Department of Communication Sciences
University of Haifa, Israel
sapir@research.haifa.ac.il

Abstract

Using two distinct data sets (from the USA and Germany) we show that machine learning tools can be used for the early diagnosis of Parkinson's disease from speech data; before physical symptoms appear. In addition, we show that while training from one country can be used in the other; different features dominate in each country; presumably because of differences in the languages. Three results are presented: (i) separate training and testing by each country (close to 85% range); (ii) pooled training and testing (80% range) and (iii) cross-country (training in one and testing in the other) (75% range). We discovered that different feature sets were needed for each country (language).

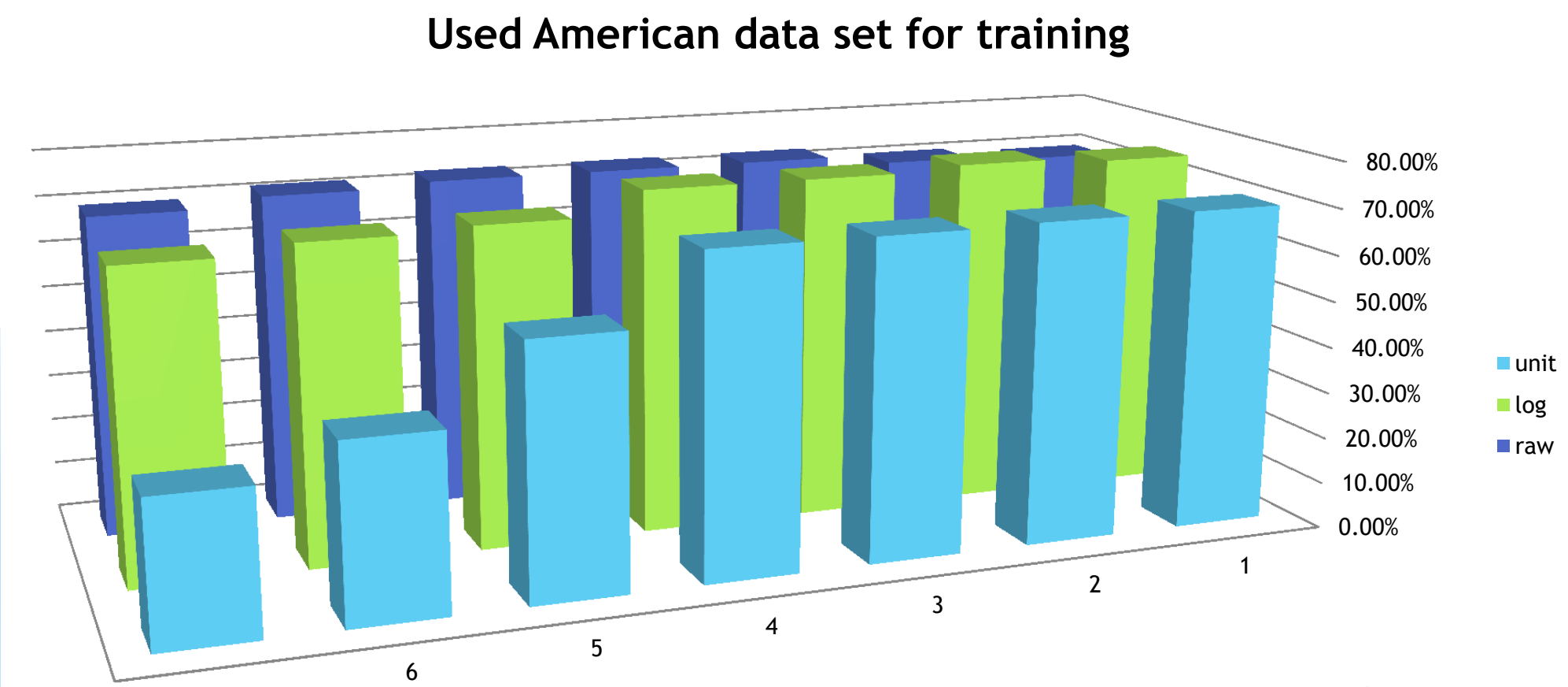
Parkinson's disease and Formants

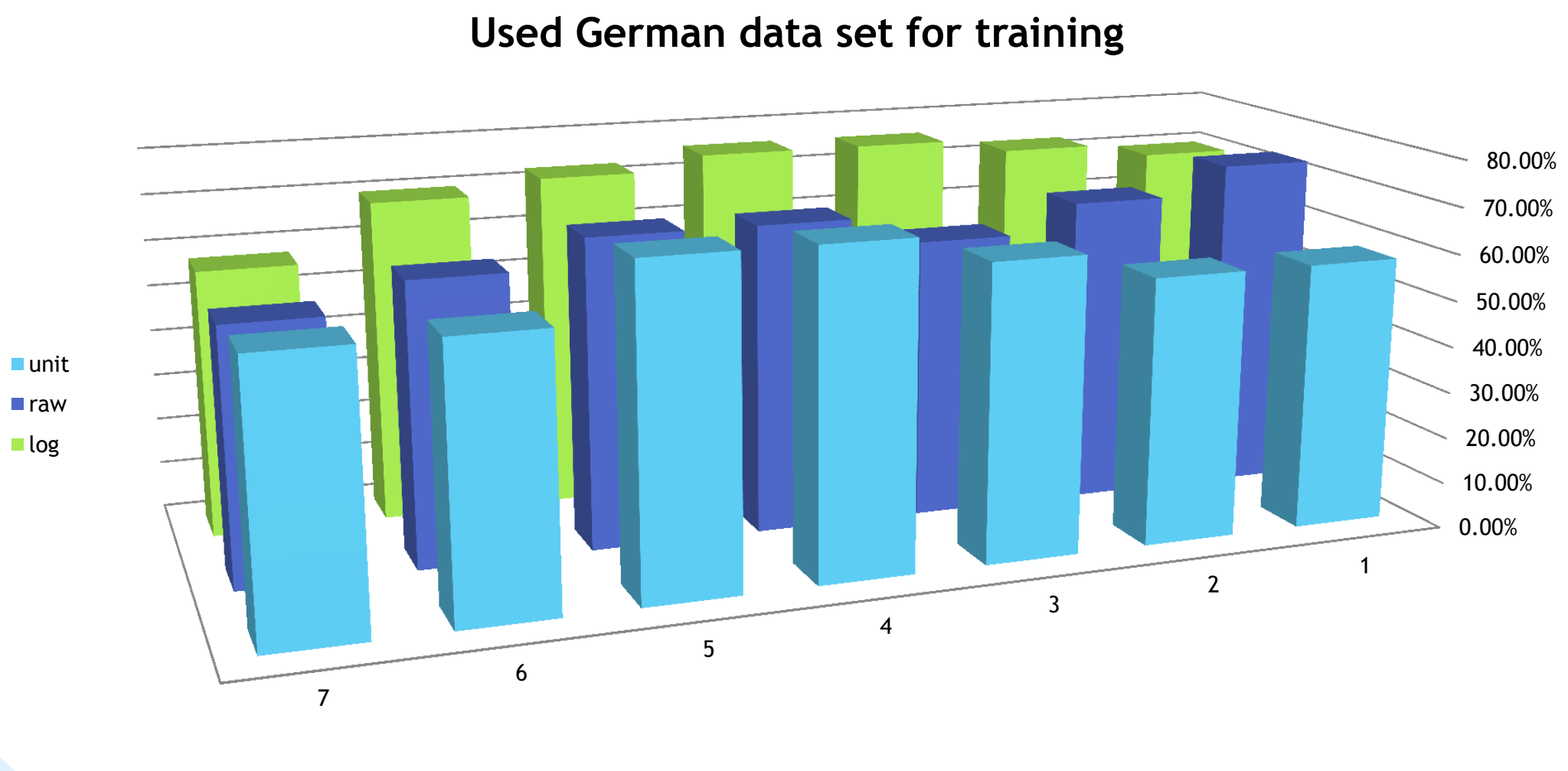
Parkinson's disease (PD) is a slowly progressing and highly debilitating disease, affecting 1.0-1.5% of individuals 60 years old and older. The disease also affects young people in their 30-50s. By the time the disease is diagnosed, some 60% of nigrostriatal neurons are degenerated, and 80% of striatal dopamine is depleted.

It has been suggested that speech abnormalities might be present at early stages of the disease, before the classical symptoms of the disease appear, and that by acoustic and classification analyses the individuals at risk for PD could be identified from the population at large.

Cross-Country

We used both of the data sets, one country's data set is used for the SVM's training process and the second country's data set is used for testing the accuracy.





The best accuracy is about 75% in each direction.

Support Vector Machine (SVM)

The Support Vectors Machine (SVM) is a tool that does a 2-class classification; that is by receiving labeled (training) data of two classes it finds an optimal hyper plane that divides the two classes. Once this has been determined, new data points are classified depending on which side of the hyper-plane they lay.

We used SVM with a radial basis function (RBF) kernel as the machine learning tool for the classification.

In this work we use machine learning on the gathered data, to classify the subjects as early Parkinson's or not.

We did the classification under three versions of representation of the data:

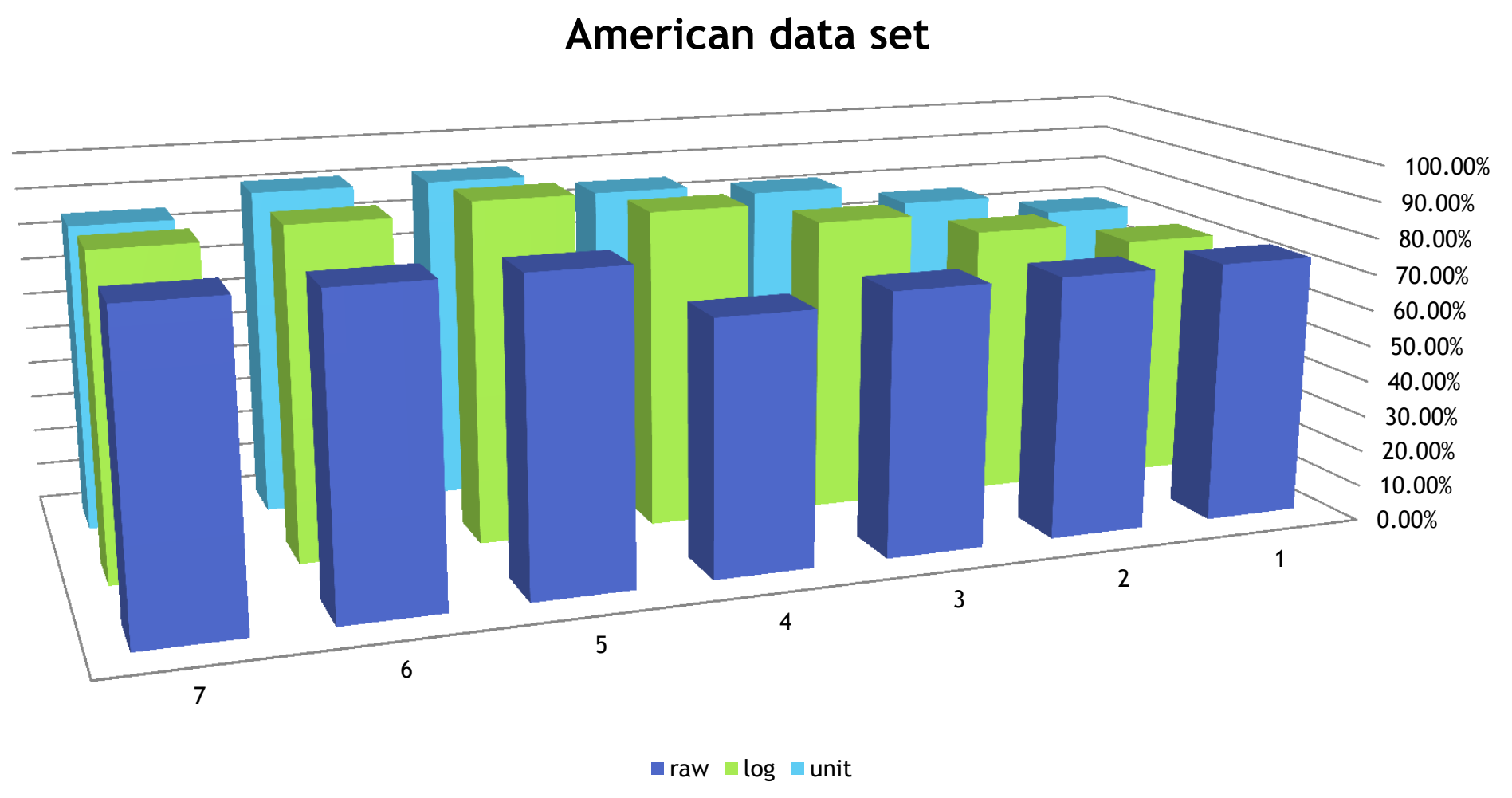
I. The raw numerical data values.
II. Unit normalization.
III. Log representation – we applied the Ln. function on each of the values.

The search was done by rerunning the SVM classification algorithm on each set of features separately.

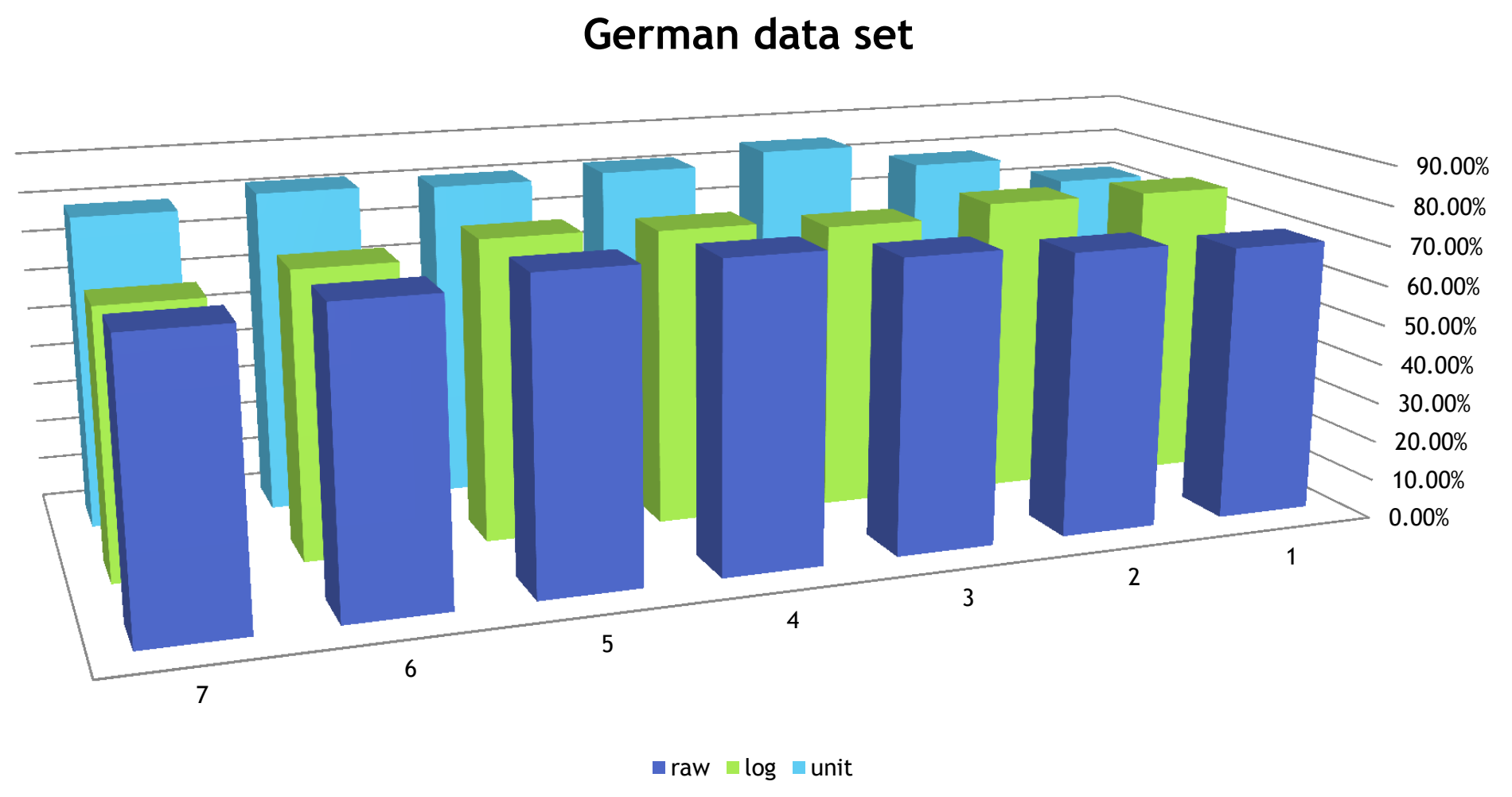
Country-wise

We examined each country's data set on its own; that is search for the best feature set that maximizes the generalization results on the country's data set using the country's data to train and to test.

We produced generalization results by using cross validation using the "leave one out" method.



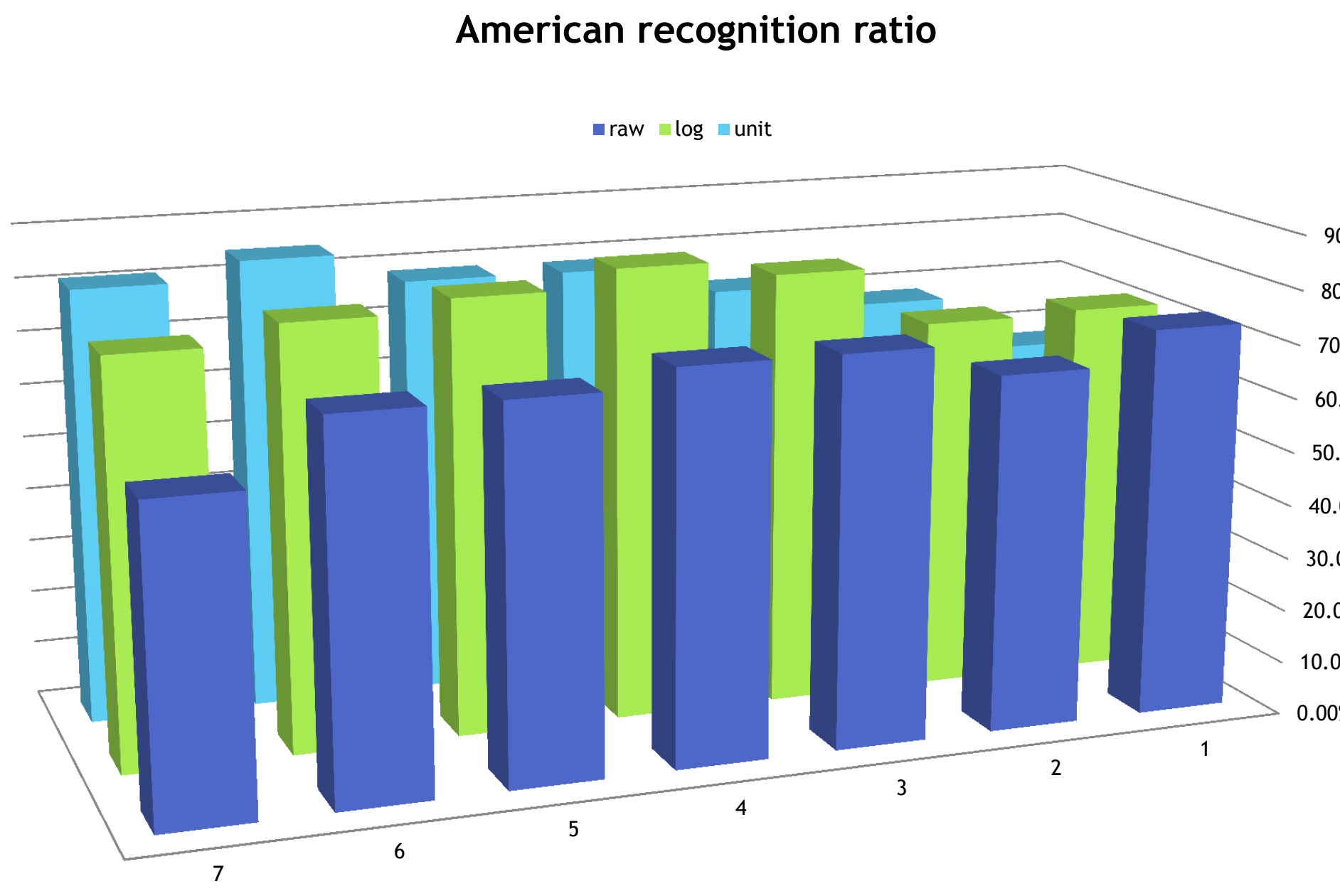
The best accuracy for the American data set is 94%.

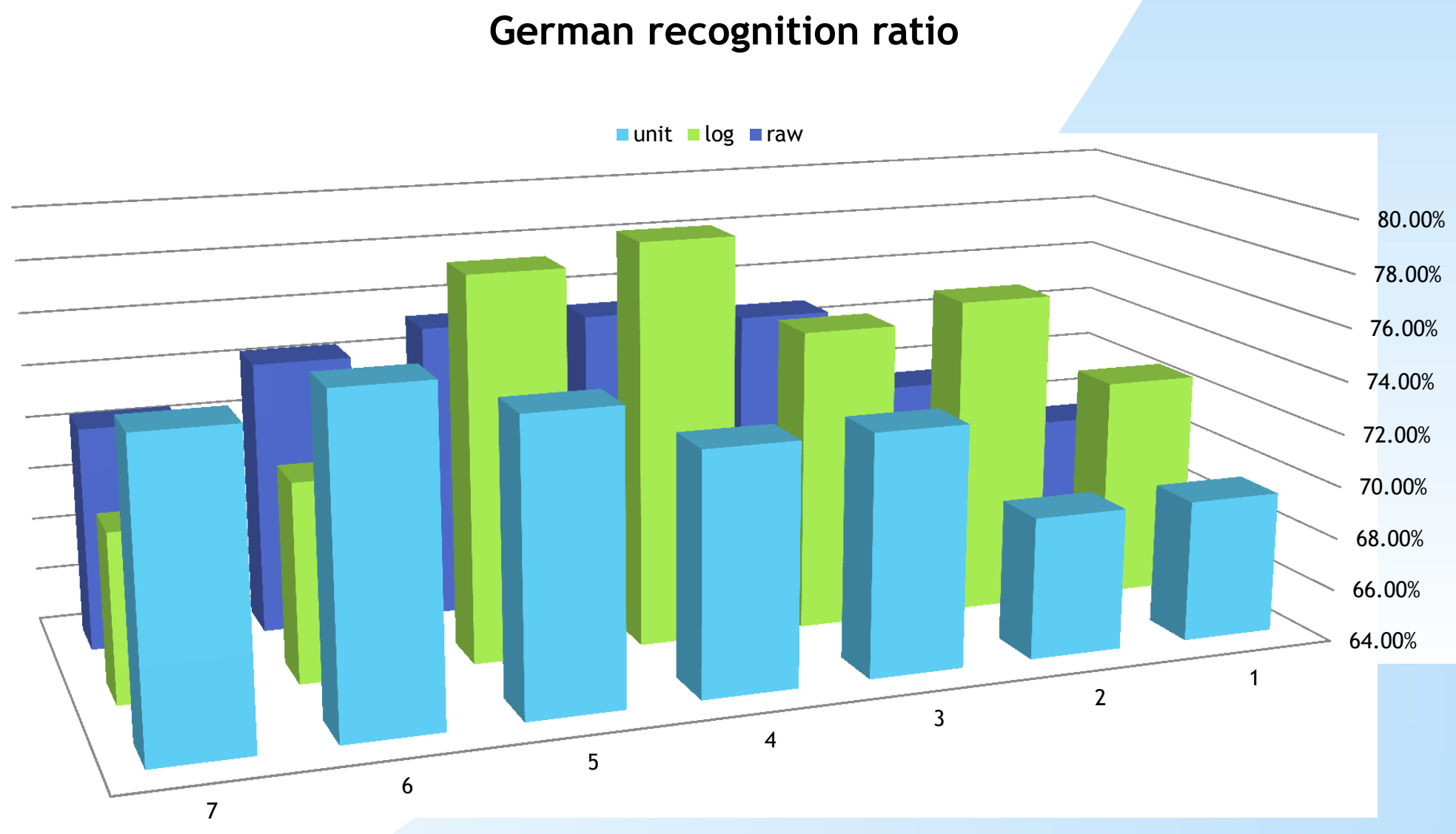


The best accuracy for the German data set is 85%.

Pooled data sets

We used the two data sets, but this time we combine the two data sets into a one large data set. The SVM trained on the large data set and we used the "leave two out" method to check how accurate was the SVMs' separation. (we took out one data point from each country's data set.)





The best result was 84% accuracy ratio for identifying the American data points and 76% accuracy ratio for identifying the German data points.

We got the best results using "**log representation**". Therefore, all of the results mentioned above were obtained using the data sets under "log representation".