Ensemble Classification for the AddNeuroMed Data Set

Dan Joyce

1 Summary

It makes intuitive sense to bolster prediction of disease (or disease state) by aggregating data from a number of different biomarkers (modalities). There are two obvious approaches (Damoulas and Girolami, 2009) shown in Figure 1, when, for a given population, participants have more than one biomarker:

1. A "brute force" approach, where all data is concatenated into one, large joint, feature-space and a single classifier is trained over this joint space
2. For each modality, train and test individual "expert" classifiers and combine their outputs to provide a final "ensemble" classification

The first "brute force" (BF) approach is straightforward to implement, but requires that all participants in the population possess all biomarkers with no missing data/samples so a robust classifier can be estimated. Performance of the BF classifier can only be evaluated on the subset of participants possessing all modality data – it cannot generally perform plausible "filling in" of missing modality data and make best-effort predictions. The latter approach provides for more flexibility, particularly in deploying the resulting classifier, when only a subset of biomarkers are available for given participant. The results presented below show that:

1. Overall, combining evidence from multiple modalities improves on the performance of individual modalities
2. Brute-force approaches – while perhaps lacking flexibility – provide performance gains over individual classifiers
3. Ensemble classifications provide a comparable level of performance to the Brute-force approach, but provide a more principled and flexible approach to the problem
4. In ensemble methods, using probabilistic methods for combining evidence appear superior to discrete voting methods

2 Data

The following data sets were extracted from the Sage databases for `AddNeuroMed`:

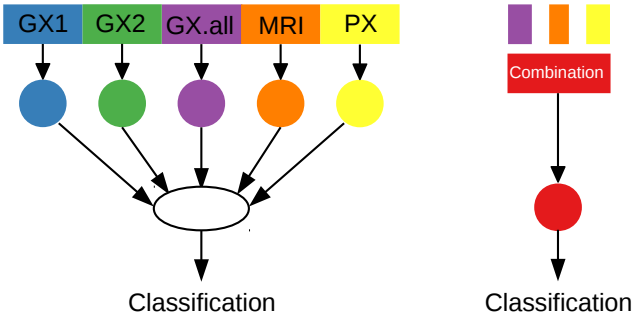

Figure 1: Evidence Combination (left); Brute-Force Combination (right); Circles indicate individual classifiers – one "expert" for each modaliity (left) and a single classifier for the combined multimodal data (right)

- Proteomics data (herein referred to as `PX`) with $n = 671$ participants and $p = 1016$ features (Sattlecker et al., 2014)
- MRI data (`MRI`) totalling $n = 163$ subjects with $p = 2150$ features from a pre-processed FreeSurfer data processing pipeline (Mangialasche et al., 2013)
- Gene expression data from (Voyle et al., 2016), where there are two 'batches' of data from two different chip arrays. Batch 1 (`GX1`) had $n = 314$ participants with $p = 6462$ features (?SNPs) and Batch 2 (`GX2`) contained $n = 250$ participants with $p = 6248$ features.
- From the two gene expression datasets (Batch 1 and Batch 2), SNPs common to both batches (resulting in $p = 5212$ common features/SNPs) were combined resulting in an aggregate (`GX.all`) set containing $n = 564$ participants. As the distributions of each feature was different between Batches 1 and 2, each was mean centred and scaled before the common features were aggregated.
- Of all participants, $n = 127$ had "complete" biomarker data for `GX.all`, `MRI` and `PX` which formed the `Combination` data set used for the brute-force approach to combining evidence.

There were some discrepancies between datasets which meant that not all participants' unique identifiers in the biomarker data could be reliably matched to clinical record data[1] and additionally, there appeared to be some duplicated samples e.g. in the proteomics data set where 6 participants with replicated samples. This resulted in modest participant loss and enabled classifiers to be built with the following sample sizes (see Figure 2):

- Proteomics (`PX`) $n = 658$
- MRI (`MRI`) $n = 163$

[1] In February 2017, the datasets were updated to include more robust cross-referencing

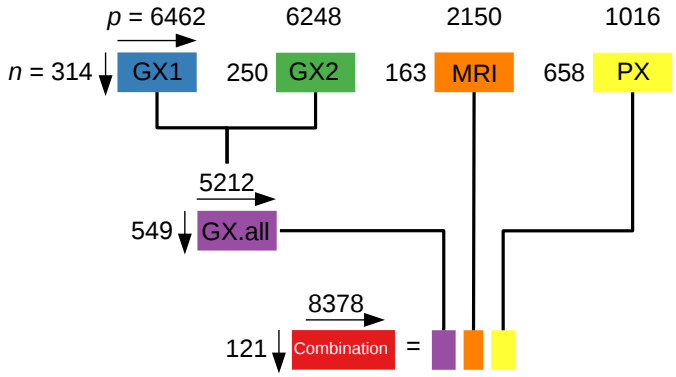

Figure 2: Sample and feature sizes for the included data sets

- Gene expression Batch 1 (`GX1`) $n = 314$
- Gene expression Batch 2 (`GX1`) $n = 250$
- Common gene expression (`GX.all`) $n = 549$
- Combination of all modalities (`Combination`) $n = 121$

3 Classifiers

Individual classifiers were built for each *modality* `MRI`, `GX1`, `GX2`, `GX.all` and `PX` – circles in the left panel of Figure 1. For each modality, samples are arranged as an $n \times p$ matrix $\mathbf{X}$, with rows X_i being a vector of individual participant's data and a corresponding target classification $Y_i \in G = \{ADC, MCI, CTL\}$ representing their assignment to either Alzheimers, mild-cognitive impairment, or controls respectively. For each modality, the aim is to model the posterior probability $\Pr(Y_i = G|X_i)$ that participant i has Alzheimers, mild-cognitive impairment or is a control participant and this is tackled as a discriminative modelling problem (Ng and Jordan, 2002). The simplest classification decision rule is the Bayes-optimal rule of assigning participant X_i to the class G with the largest posterior probability (Bishop, 2007).

Given that the number of features/predictors greatly exceeds the number of participants ($p \gg n$) in each modality, regularised multinomial generalised linear models (so-called *elastic-net* GLMs) were used (Friedman et al., 2010) to find an optimal and sparse set of predictors for the final classifier. The elastic-net implements a continuum between the lasso- and ridge-regression penalty solutions. The former tends to retain the largest, non-zero coefficients for predictors (discarding the remaining features) and the latter shrinks estimated coefficients of correlated (redundant) predictor variables toward each other. The lasso penalty generally results in smaller, sparser sets of features in the final model (Friedman et al., 2010). For the classifiers implemented here, the lasso penalty was used. There is no in-principle reason why other classifier algorithms cannot be

used – e.g. Gaussian process classifiers (GPC), support vector machines (SVM) or orthogonal partial least squares to latent structures (OPLS). A robust, but computationally expensive training method was used to minimise classifier bias (underfitting) and prevent high variance (overfitting) – see (Breiman et al., 1996), (Friedman et al., 2001) and (Bengio and Grandvalet, 2004). Therefore, a relatively fast, maximum likelihood-based method was used (versus for example, the more computationally costly Bayesian estimation of GPCs). A further requirement for evidence combination (Figure 1, left panel) is that the classifier can naturally yield estimates of the probabilities $\Pr(Y_i = G|X_i)$ – which is more difficult when using e.g. SVMs and OPLS because of their "one versus all" approach to multi-class problems.

4 Classifier Training

To train a classifier for a modality requires a regularisation parameter, λ, which sets a threshold for excluding predictors that are not contributing to classification. This model-selection process finds the optimal value of λ – where each value yields a particular solution and classifier given the data – and the bias of the classifier is minimised. While cross-validation can be used to select the optimal λ, this requires further embedding in a model assessment stage that guards against fortuitous choices for the fold partitioning (e.g. high variance solutions). Classifier training, model selection (for parameter λ) and model assessment (for out-of-sample classification performance) was performed using repeated, nested, stratified k-folds cross-validation. This differs from the traditional 10-fold stratified method (Kohavi, 1995) often used for model selection, and instead implements the algorithm given in the Appendix, (adapted from Algorithm 2 in (Krstajic et al., 2014) for the GLMnet classifier).

For the results presented here, the number of inner and outer loops was $N_1 = N_2 = 50$ (essentially, the number of times cross-validation for model selection and assessment are repeated) and the number of inner and outer folds was $V_1 = V_2 = 10$. The advantage of this method is that:

- Estimates of classifier performance – i.e. an estimate of $\Pr(Y_i = G|X_i)$ for every participant in the data set – is an average of N_2 repetitions of V_2-fold cross-validation (the outer loop)
- The optimal regularisation parameter λ is estimated by averaging over N_1 repeated inner loops of V_1-cross-validations
- Optimistic estimates of $\Pr(Y_i = G|X_i)$ due to "lucky" partitioning of the data into V_2 folds is avoided by repeating the outer loop N_2 times
- For each participant i, the classifier used to estimate the probability $\Pr(Y_i = G|X_i)$ was trained *only* on a subset of participants[2] which *did not* include i

[2]In all the following results, performance is reported *only* using out-of-sample estimates for posterior probabilities so classification results are never over-optimistic. Note, unlike some published results, the model parameters obtained from cross-validation model selection are not then used to re-train a model with *all* available data which artificially boosts reported classifier accuracy

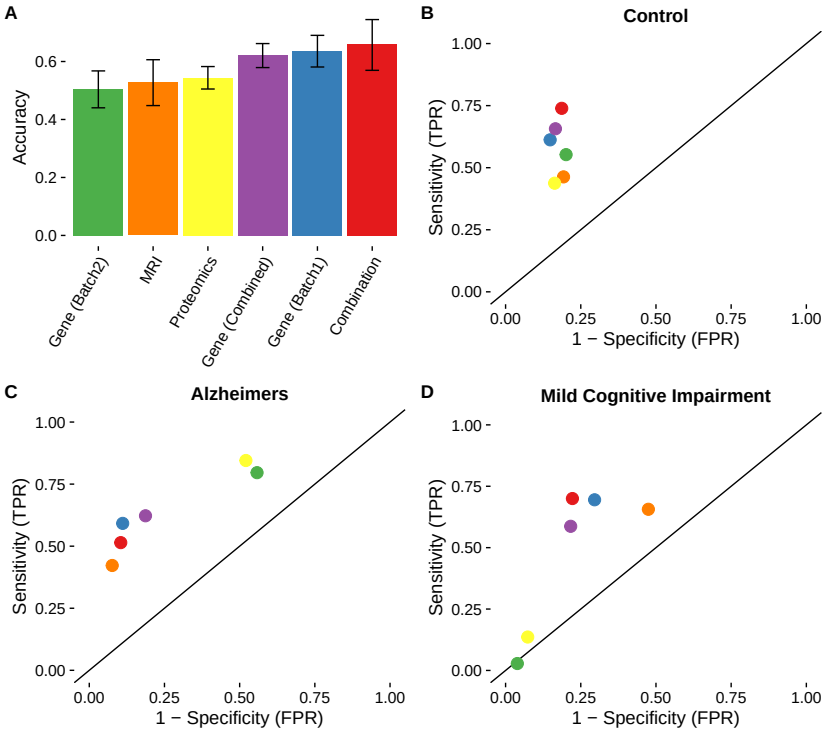

Figure 3: For individual modalities, the complete data sets were used (`PX`, $n = 658$; `MRI`, $n = 163$; `GX1`, $n = 314$; `GX1`, $n = 250$; `GX.all`, $n = 549$); For the "brute force" classifier, `Combination`, $n = 121$. A: Overall Accuracy (correct versus incorrect) for each dataset / modality; ROC space of modality-classifiers for each diagnostic class B: Classifying Controls correctly (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars in A represent 95% confidence intervals for a Binomial test of accuracy significantly different from the 'no information rate' (NIR), which is taken to be the class with the highest prevalence; i.e. if different diagnoses are represented equally in the data set, then the NIR = 1/3

5 Individual Modality and Brute-Force Classifier Performance

To test *individual* classifier's performance, for a given participant their predicted class from a modality was simply "hard assigned" on the basis of the highest posterior probability (i.e. the simple Bayes-optimal decision rule). For example, if a classifier reports for a given participant the posterior probabilities:

$$\Pr(Y_i = ADC|X_i) = 0.35$$
$$\Pr(Y_i = MCI|X_i) = 0.36$$
$$\Pr(Y_i = CTL|X_i) = 0.29$$

Then the participant is predicted to be in the MCI (mild cognitive impairment) class. In terms of assessing prediction error, a 0-1 loss function is assumed such that if the above participant actually has a diagnosis of MCI then the classifier scores one for this example (conversely, if the predicted class was ADC or CTL, then the score is zero). This affords an intuitive overall "correct versus incorrect" accuracy score, but is less informative than the true- and false positive rates for a given diagnosis (i.e. the classifier's performance in ROC space).

Figure 3 shows the performance of 5 classifiers (trained as described above) for each data set `PX`, `MRI`, `GX1`, `GX2` and `GX.all` – i.e. the performance for the individual classifiers in Figure 1 (left). In addition, the 'brute force' combination classifier was implemented (shown in red, corresponding to Figure 1, right panel), trained on the $n = 121$ participants with complete data for `PX`, `MRI`, and `GX.all` as described in Figure 2.

Overall, the 'brute-force' (BF) combination classifier performs favourably over individual modalities alone in predicting classification; however, classifier performance can only be evaluated on the small ($n = 121$) subset of participants who possess data on all modalities; this issue is taken up later.

6 Challenges for Evidence Combination

Using the BF method (Figure 1, right), classification is performed on a concatenated feature space of 8378 features. There are two significant difficulties presented by this approach:

1. If, for example, a sub-group of participants do not have e.g. proteomics data, then for these participants, around 12% of their data are systematically missing. Similarly, in another sub-group with gene expression and proteomics, but no MRI data, then for these participants, around 25% of their data is systematically missing.

2. The BF method requires one concatenated feature-space to be formed and, if each modality has different distributions, the classifier is forced to cope with heterogeneity in distributions of this combined feature space – an extreme example being augmenting two modality's feature spaces, one with binary features and the other with continuous, normally distributed features.

In a situation where a participant has, for example, only a proteomics biomarker, obtaining a classification using the brute-force approach is impossible; the single classifier has been trained on a concatenated feature space. A plausible model of missing data is difficult to implement, unless a generative model of the joint distribution of $(\mathbf{X}, Y)$ is available during classifier construction (generally, this is intractable).

The more statistically concerning problem with brute-force classification is illustrated in Figure 4, where two simulated, unidimensional modalities ($p = 1$) are shown for $n = 1000$. In each modality `X1` and `X2`, the true classification is given by a threshold[3]. Classification within each modality is trivial – if a participant has `X1` < 7 they are assigned Class 0, otherwise Class 1. Similarly,

[3]This toy example is used to illustrate the problem; it would be rare for a discrete, decision threshold to exist which unambiguously defines class membership

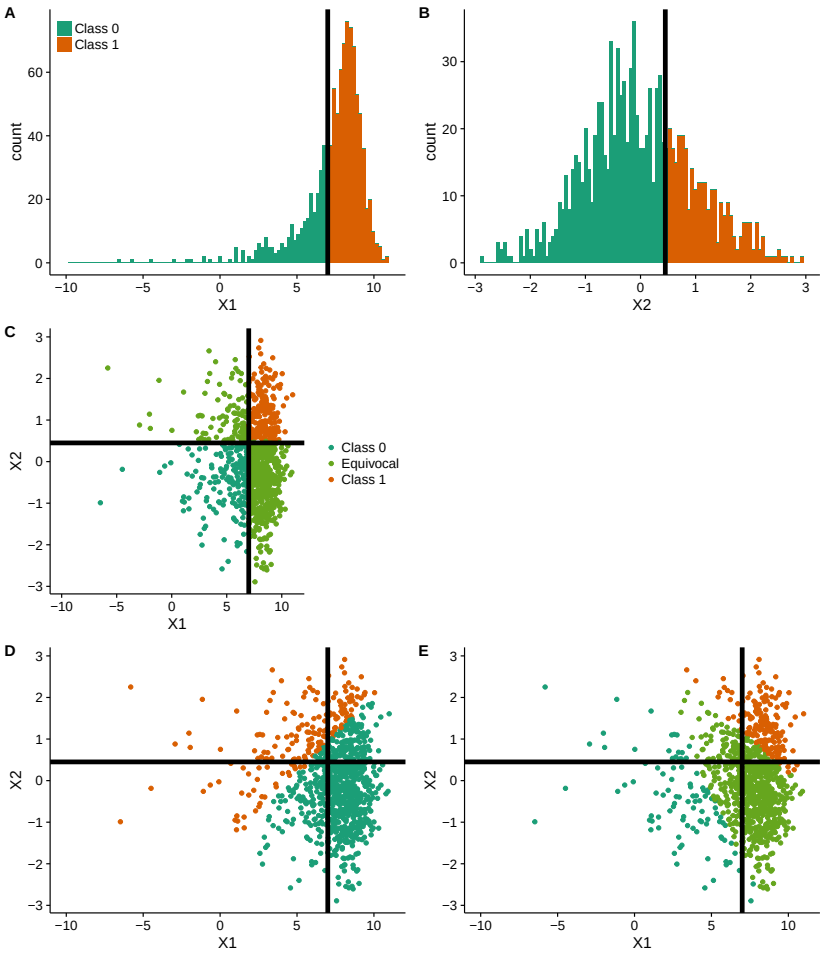

Figure 4: Hypothetical Classification Problem – A: Individual modality `X1` with a unidimensional "feature-space", and a trivial classification into Classes 0 and 1; B: Similarly, for modality `X2`; Classification in the native feature space (modality) is straightforward and unequivocal. C: Using the brute-force approach, a joint feature-space is formed by concatenating `X1` and `X2`; in this concatenated feature-space, a number of participants are now in an *equivocal* class where the assignments in `X1` and `X2` disagree; D: GLM attempting to classify using the original two classes in the concatenated feature space; E: classification using a more flexible multinomial GLM allowing for an additional class when there is equivocation (disagreement) resulting from concatenating the two modality's native feature spaces.

in the second modality, if a participant has `X2` < 0.5, they are assigned Class 0, otherwise Class 1.

The first difficulty is illustrated in Figure 4, panel C: the concatenated, joint feature space (`X1`, `X2`) results in a situation where some samples in `X1` have different classifications in `X2`, so the overall classification is equivocal. Using this concatenated feature space, with a generalised linear model (GLM) over

the two classes results in Figure 4, panel D with poor classification. Using a more flexible multinomial GLM gives the result shown in Figure 4, panel E. While this is more robust, the brute-force approach requires accounting for disagreement in the concatenated 2D feature space. While there are solutions to these problems, they will be specific to individual datasets; for example, one could exploit covariance structure in the concatenated feature space and use kernel-based methods see e.g. (Damoulas and Girolami, 2009).

An alternative, flexible proposal is to use evidence combination over the posterior probabilities returned by each individual modality (Figure 1, left). This allows for "best attempt" classification when participants only have one or two biomarkers (instead of the full complement required by the 'brute-force' method) and allows for classifiers to be "modality experts", learning properties of the conditional distribution for their respective modalities.

7 Approaches to Combining Evidence from Individual Modality Classifiers

For a review of ensemble-based classification systems, see (Polikar, 2006) and (Kittler, 1998) for a more detailed probabilistic framework and theoretical assumptions required for the probabilistic rules adopted below. To make terminology consistent, the aggregation of outputs from individual, modality-specific classifiers is termed *ensemble classification*. The following notation is used:

- there are classifiers for each modality $M = \{\texttt{PX}, \texttt{MRI}, \texttt{GX1}, \texttt{GX2}, \texttt{GX.all}\}$
- as before, there are three classes $G = \{ADC, CTL, MCI\}$
- the posterior probability of participant X_i being in a class G, reported by a particular modality classifier is $\Pr(G|X_i; m_j)$ where $j \in M$
- the *discrete* classification – or 'vote' – from individual modality classifier j is the class G with the largest posterior probability:

$$v_j = \underset{G}{\operatorname{argmax}} \Pr(G|X_i; m_j)$$

As an example, assume the modality classifiers (circles in Figure 1, left) report posterior probabilities as shown in Table 1. Methods for ensemble classification over the posteriors are:

- **discrete** voting methods – the final classification is given by taking the class G with maximum number of votes v_j
- **heuristic** methods – where probabilities reported by each modality are combined by e.g.averaging where the **mean** or **median** of the columns in Table 1 are computed, and the class G with the largest support is chosen
- **probabilistic** methods – where under defined independence assumptions for modality classifiers, the **product** or **sum** of the posterior probabilities are computed
- **meta-learning** – where classifier combinations are learned using "stacking" or "boosting" methods (see Future Work discussion below).

m_j	$\Pr(ADC\|X_i)$	$\Pr(CTL\|X_i)$	$\Pr(MCI\|X_i)$	v_j
`MRI`	0.08	0.19	0.73	MCI
`PX`	0.99	0.01	0.00	ADC
`GX1`	0.01	0.00	0.99	MCI
`GX2`	0.41	0.30	0.28	ADC
`GXall`	0.02	0.72	0.26	CTL
		Ensemble Combination		Classification
Votes				MCI or ADC
Mean	0.30	0.24	0.45	MCI
Median	0.08	0.20	0.28	MCI
Product	0.8×10^{-5}	0.0	0.0	ADC
Sum	0.20	0.21	0.58	MCI

Table 1: Example posterior probabilities returned by individual modality classifiers in the ensemble shown in Figure 1 (left) with different combination strategies shown below

Note that in all cases, results presented for individual, ensemble and BF classification performance are obtained *only* using out-of-sample estimates from the model-selection/assessment procedure.

7.1 Evidence Combination Compared to Brute-Force Methods

Using only participants with all modalities ($n = 121$), the BF method can be compared with the evidence-combination methods described above. Figure 5 shows that discrete, voting methods do not perform favourably but there are modest gains obtained using the mean, product and sum rules. The ROC space reveals that ensemble methods generally produce favourable results when compared with the BF method (illustrated by bootstrapped mean false- and true-positive rates being shifted up and toward the left corner of ROC space). While these performance gains are modest, the flexibility afforded can be exploited when – as described below – attempts are made with less-than full modality data on the complete available data.

7.2 Ensemble Classification: Effect of Increasing Available Modalities

In comparing ensemble methods with the BF method so far, only the restricted data set of $n = 121$ participants with all biomarkers has been used. Here, flexibility is demonstrated when the ensemble approach is applied to subsets of participants with incremental numbers of modalities. In situations where a participant lacks a specific modality, the responsible classifier reports the population, modality-specific prior probability $\Pr(G; m_j)$. Essentially, classifiers report there estimate of $\Pr(G|X_i; m_j)$ and when data is not available, the classifier j reports only it's prior. The pool of all available data is $N = 874$ participants with between 1 and 4 modalities available, corresponding to `MRI`,

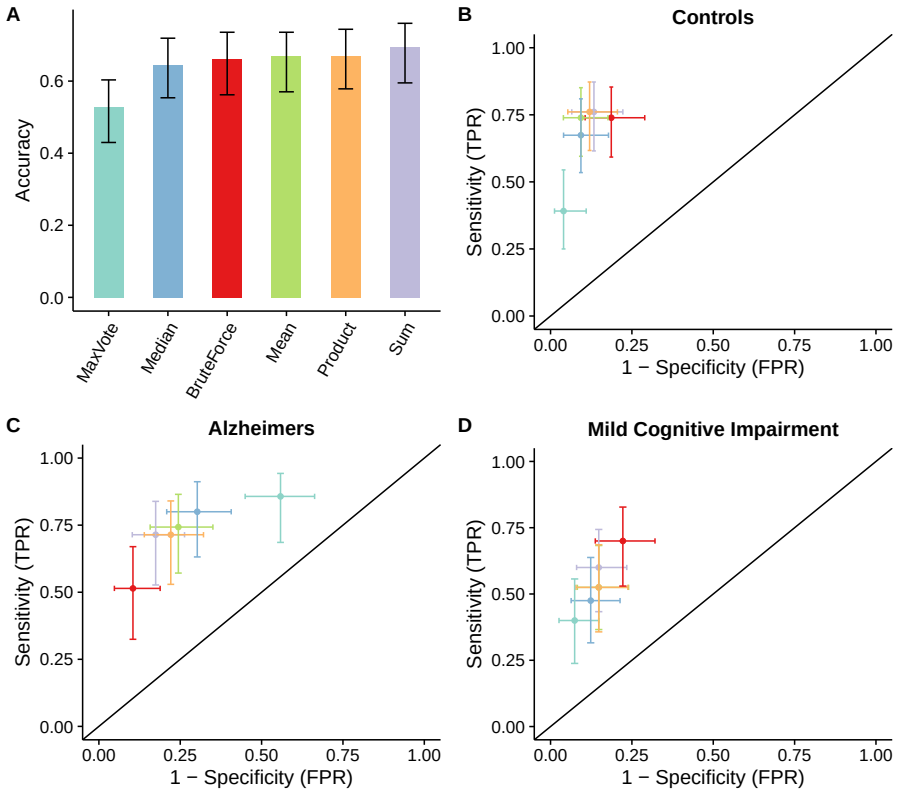

Figure 5: A: Overall Accuracy (correct versus incorrect) for $n = 121$ participants with data in all modalities; ROC space of the different evidence combination classifiers for each diagnostic class B: Classifying Controls (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars represent bootstrapped 95% confidence intervals on performance statistics

`PX`, `GX1` or `GX2` and `GXall`. Participants have data in either `GX1` or `GX2` and data in the combined gene expression data set `GXall`, hence the largest number of modalities a participant can have is four.

Algorithm 1 describes how the experiments were run. In summary, sub-populations of participants are selected who have one modality (irrespective of *which* modality) and bootstrapping is used on the subset to obtain ROC statistics for classification performance. This is repeated with the sub-population having two, then three and finally, four modalities. These results describe how adding modalities change classification performance.

Figures 6, 7 and 8 show ensemble classification performance as a function of the number of available modalities for the heuristic mean, probabilistic product and sum rules respectively[4].

For all ensemble rules, having more modalities increases performance, with little gain moving from 3 to 4 modalities. In terms of the optimal combination

[4]The top three methods from the previous experiments were used – see Figure 5

Algorithm 1: Computing Performance of Ensemble Classifications as a Function of the Number of Modalities

Let $m = \{1, 2, 3, 4\}$ be the number of modalities
Let S_m be the subset of $N = 874$ participants possessing a total of m modalities
for *2000 times* **do**
  Let $S_m^{'}$ be a bootstrap (with replacement) resample of S_m
  forall *participants* X_i *in* $S_m^{'}$ **do**
    Obtain posterior probabilities $\Pr(G|X_i; m_j)$ by submitting participant data to individual modality classifiers
    Combine posterior probabilities and classify X_i
  end
  Compute ROC performance statistics (TPR, FPR, Accuracy) for this bootstrap
end
Result: Bootstrapped ROC performance estimates and 95% confidence intervals for participants having at exactly m modalities

rule, the mean rule outperforms the probabilistic product rule for true- and false positive rates but MCI and Alzheimer's disease classification is marginally better with the probabilistic sum rule (i.e. lower false positive rates with comparable true-positives). There is little to recommend one rule over the others, with perhaps the exception that the sum rule appears to control false positive rates over all three diagnostic classes more consistently than the product or mean rules. Theoretically, the product and sum rules are more justified in their development and assumptions with the mean rule being heuristic.

7.3 Which Combinations Perform Best?

In the preceding section, the effect of adding modalities was compared with different ensemble rules, but no weight was given to *which* modalities where driving classification performance. Here, the sub-populations were selected by sequentially testing each combination of one, two or three modalities. For example, participants with `MRI` *and* `PX` can be compared against the sub-population of participants with *only* `MRI` and *only* PX alone. Similarly, combinations such as `PX` *and* `GX` versus those with only `MRI`. Naturally, because of the design of the study, some combinations and single modalities have very small sample sizes – for example, 7 participants had only `MRI` *and* `PX`. To reduce the number of combinations to test, for the `GX` samples, participants were selected from the combined `GXall` modality rather than individual batches `GX1` and `GX2`.

Figure 9 shows how different modalities in isolation or combined change classification performance using the sum rule (chosen for its theoretical and empirical properties, described above). The best performance is obtained when combining `MRI`, `PX` and `GX` and this is pronounced when examining the ROC statistics for each diagnostic class. The sub-population combining `PX` and `GX` performs favourably but marginally sacrifices performance on false positive rates for Alzheimer's and Mild Cognitive Impairment. Conversely, these results can

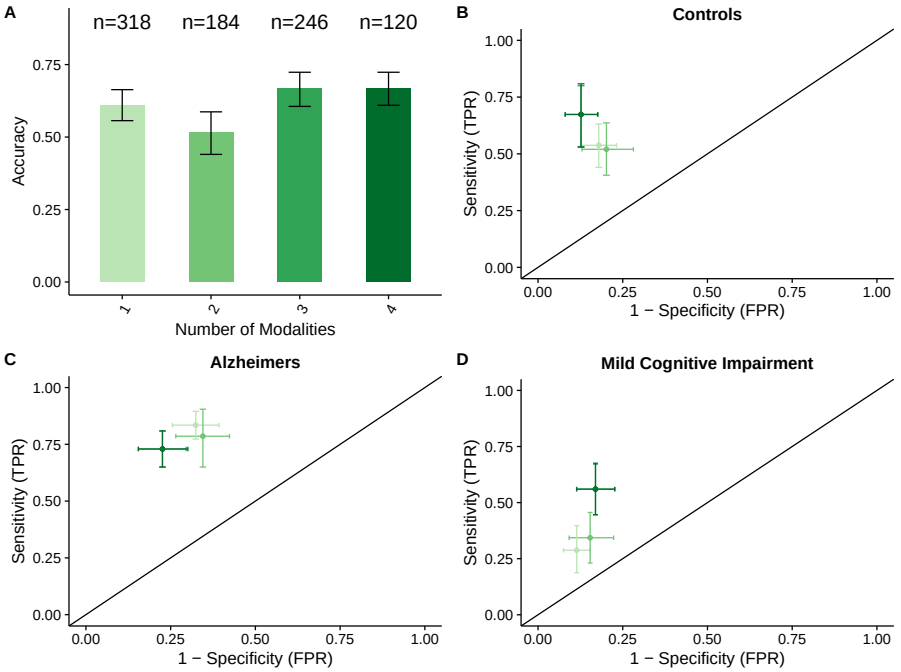

Figure 6: Ensemble Classification using the Mean Rule – A: Overall Accuracy (correct versus incorrect) for all participants; ROC space of the different evidence combination classifiers for each diagnostic class B: Classifying Controls (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars represent bootstrapped 95% confidence intervals

be interpreted as suggesting combining proteomics with gene expression data gives a baseline ensemble performance that improves slightly – for diagnosing Alzheimer's and Mild Cognitive Impairment – when neuroimaging data is added.

8 Future Work

1. **Outstanding Inferential Statistics for Performance**: given that performance over individual classes (e.g. ROC statistics presented in Figures 5 – 9) is relevant in this data, a workable inferential method for determining superiority is required (e.g. beyond simple tests on global accuracy).

2. **Learning Combinations**: The above results illustrate evidence combination by ensembles using *fixed* functions. The next step is to try so-called meta-learning methods that attempt to learn the best ensemble – for example, variants on "stacking" (Wolpert, 1992; Breiman, 1996; Sill et al., 2009) and Bayesian model combination (Monteith et al., 2011).

3. **Individual Classifiers**: The GLMnet algorithm is not necessarily the optimal classifier algorithm for each modality – for example, gene expression data may be better classified using shrunken centroid-based methods

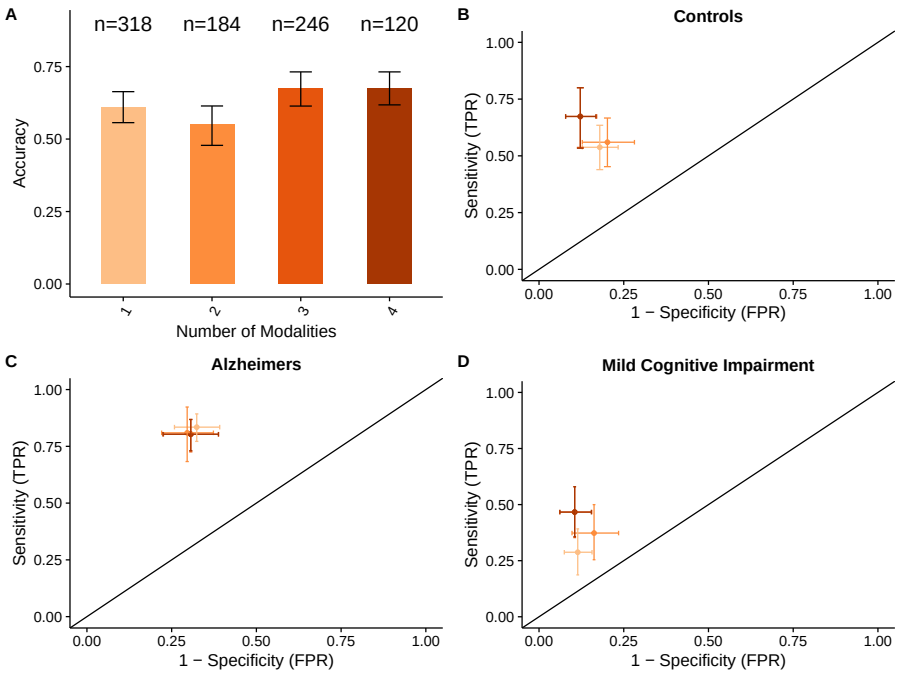

Figure 7: Ensemble Classification using the Product Rule – A: Overall Accuracy (correct versus incorrect) for all participants; ROC space of the different evidence combination classifiers for each diagnostic class B: Classifying Controls (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars represent bootstrapped 95% confidence intervals

(Tibshirani et al., 2002), although initial results were not impressive on this dataset. Of note, having many classifiers each using different algorithms on the same modality is easily provided for in the ensemble methods framework.

4. **Representational issues**: features are currently entered "raw" into classifiers and there is evident redundancy as the GLMnets find a sparse feature set to classify. If there is a parsimonious topology underlying the features - this could be found by unsupervised methods first e.g. locally-linear embedding (Roweis and Saul, 2000), isomaps (Tenenbaum et al., 2000) or simply distance-preserving multidimensional scaling. The former remain controversial (Balasubramanian and Schwartz, 2002) and require adaptations for use in classification where training/testing relies on out-of-sample validation (Bengio et al., 2003). Reduced-dimensional representations could be properly explored for their potential to enhance both modality-level classifier performance, and improve combination - whether by brute force, or by evidence combination. Further so-called multi-kernel methods may offer a unified approach to brute-force methods (Damoulas and Girolami, 2009).

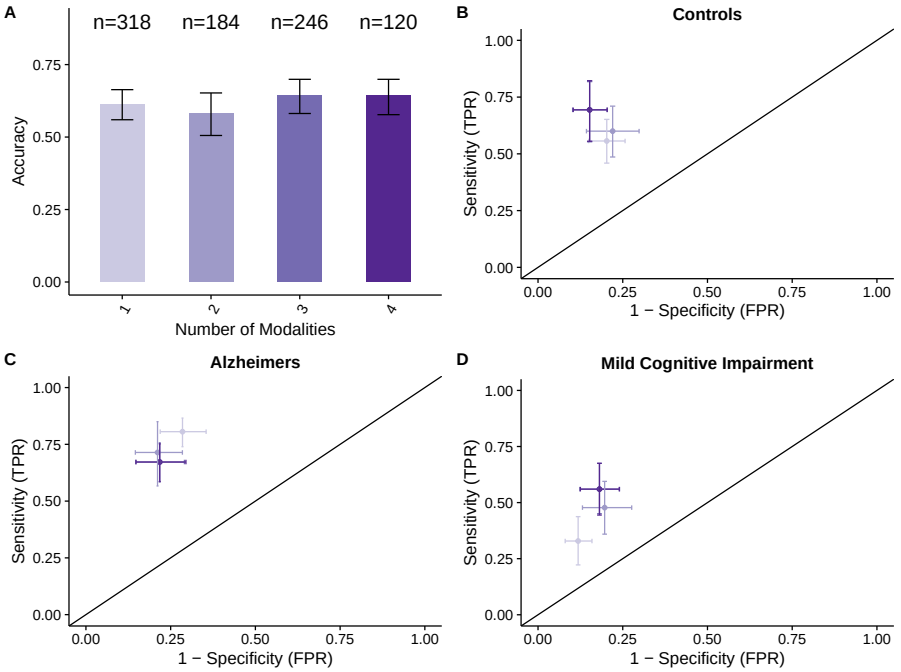

Figure 8: Ensemble Classification using the Sum Rule – A: Overall Accuracy (correct versus incorrect) for all participants; ROC space of the different evidence combination classifiers for each diagnostic class B: Classifying Controls (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars represent bootstrapped 95% confidence intervals

5. A reasonable question is **"Why not DeepLearning?"**: given the suggestion of exploiting data dimensionality-reduction (representation), as well as the simplicity of evidence combination using brute-force methods, hierarchical supervised learning of increasingly sparse representation at the same time as classification has appeal; for example, as in convolutional networks (Krizhevsky et al., 2012). There are difficulties understanding *what* these networks learn and the interpretation of the outputs – while probabilistic – are not obviously understood in the same way as e.g. GLM-based models that preserve the interpretation of each classifier as a discriminative model with outputs readily interpreted as posterior probabilities of class membership (under the rubric of statistical decision theory). This remains an active area of theoretical and empirical work, but for diagnostic decision making, it may be prudent to retain the more traditional interpretation afforded by well-understood methods.

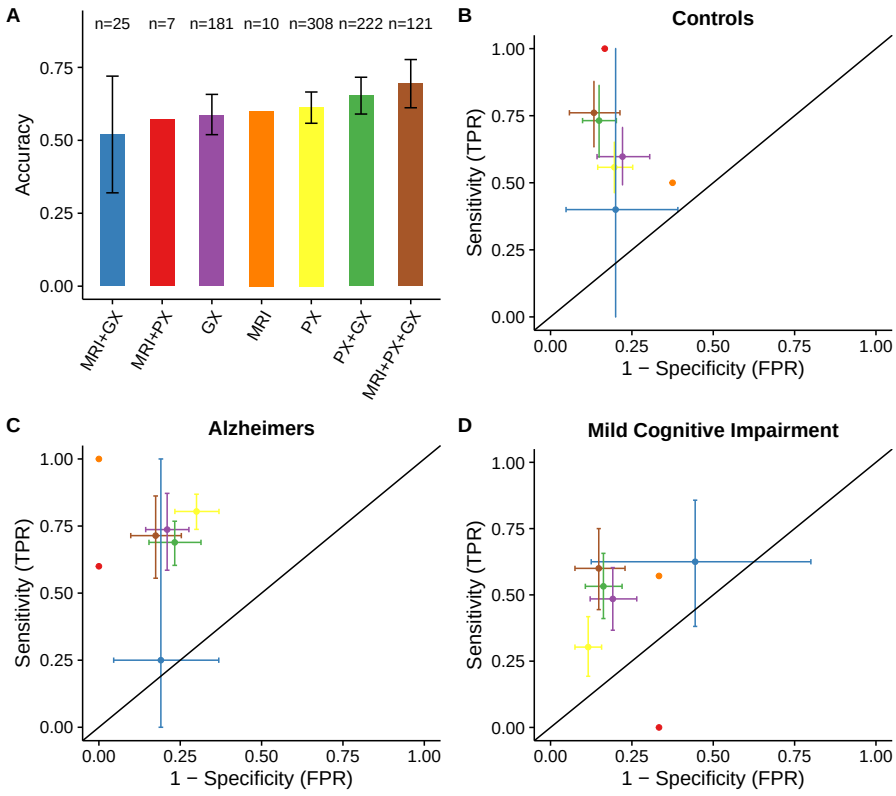

Figure 9: Comparitive Ensemble Classification Performance by Modality (using Sum Rule) – A: Overall Accuracy (correct versus incorrect); ROC space of the different evidence modalities classifiers for each diagnostic class B: Classifying Controls (versus both other groups); C: Alzheimer's (versus others) and D: Mild Cognitive Impairment (versus others). Error-bars represent bootstrapped 95% confidence intervals; For `MRI+PX` and `MRI` alone, n was too small to estimate robust confidence intervals

A Repeated, Nested Cross-Validation

Algorithm 2: Repeated, nested cross-validation

for N_2 *times* **do**
  Stratify the entire data set $D = (\mathbf{X}, Y)$ into V_2 pseudo-random folds with balanced proportions of classes G in each fold
  forall *folds* i *in* V_2 **do**
    Let this training set, τ_2, be D *excluding* fold i
    Let this validation set, ν_2, be fold i
    Model Selection:
    for N_1 *times* **do**
      Divide τ_2 into V_1 folds
      forall *folds* j *in* V_1 **do**
        Let this inner training set, τ_1 be τ_2 *excluding* fold j
        Let this inner validation set ν_1 be fold j
        Search for optimal λ_j by training a classifier F on τ_1 that maximises classification performance on ν_1
      end
      Store each λ_j value for each V_1 fold
    end
    Compute final $\hat{\lambda}$ as the average of the N_1 values of λ_j
    Model Assessment:
    Build a final classifier F_i with $\hat{\lambda}$ on training set τ_2
    Use F_i to predict $\Pr(Y = G|X)$ for participants in validation set ν_2
  end
end
Result: N_2 estimates of $\Pr(Y = G|X)$, obtained out-of-sample, for each participant in D

Final Model:
Compute and store average of N_2 out-of-sample estimates of the probability $\Pr(Y = G|X)$ for each participant

References

Balasubramanian, M. and Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, 295(5552):7–7.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Roux, N. L., and Ouimet, M. (2003). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 177–184. MIT Press.

Bishop, C. (2007). Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*.

Breiman, L. (1996). Stacked regressions. *Machine learning*, 24(1):49–64.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Damoulas, T. and Girolami, M. A. (2009). Combining feature spaces for classification. *Pattern Recognition*, 42(11):2671–2683.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Kittler, J. (1998). Combining classifiers: A theoretical framework. *Pattern analysis and Applications*, 1(1):18–27.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143, San Francisco. Morgan Kaufmann Publishers.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10.

Mangialasche, F., Westman, E., Kivipelto, M., Muehlboeck, J.-S., Cecchetti, R., Baglioni, M., Tarducci, R., Gobbi, G., Floridi, P., Soininen, H., et al. (2013). Classification and prediction of clinical diagnosis of alzheimer's disease based on mri and plasma measures of α-/γ-tocotrienols and γ-tocopherol. *Journal of internal medicine*, 273(6):602–621.

Monteith, K., Carroll, J. L., Seppi, K., and Martinez, T. (2011). Turning bayesian model averaging into bayesian model combination. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2657–2663. IEEE.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 2:841–848.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Sattlecker, M., Kiddle, S. J., Newhouse, S., Proitsi, P., Nelson, S., Williams, S., Johnston, C., Killick, R., Simmons, A., Westman, E., et al. (2014). Alzheimer's disease biomarker discovery using somascan multiplexed protein technology. *Alzheimer's & Dementia*, 10(6):724–734.

Sill, J., Takács, G., Mackey, L., and Lin, D. (2009). Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Voyle, N., Keohane, A., Newhouse, S., Lunnon, K., Johnston, C., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., et al. (2016). A pathway based classification method for analyzing gene expression for alzheimers disease diagnosis. *Journal of Alzheimer's Disease*, 49(3):659–669.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.