Prediction of Motor Function in Very Preterm Infants using Connectome Features and Local Synthetic Instances

Colin J Brown[1], Steven P Miller[2], Brian G Booth[1], Kenneth J Poskitt[3],
Vann Chau[2], Anne R Synnes[3], Jill G Zwicker[3], Ruth E Grunau[3]
and Ghassan Hamarneh[1]

[1]Simon Fraser University, BC, Canada, [2]The Hospital for Sick Children and The
University of Toronto, ON, Canada, [3]University of British Columbia and Child and
Family Research Institute, BC, Canada

Abstract. We propose a method to identify preterm infants at highest
risk of adverse motor function (identified at 18 months of age) using con-
nectome features from a diffusion tensor image (DTI) acquired shortly
after birth. For each full-brain DTI, a connectome is constructed and net-
work features are extracted. After further reducing the dimensionality of
the feature vector via PCA, SVM is used to discriminate between normal
and abnormal motor scores. We further introduce a novel method to pro-
duce realistic synthetic training data in order to reduce the effects of class
imbalance. Our method is tested on a dataset of 168 DTIs of 115 very
preterm infants, scanned between 27 and 45 weeks post-menstrual age.
We show that using our synthesized training data can consistently im-
prove classification accuracy while setting a baseline for this challenging
prediction problem. This work presents the first image analysis approach
to predicting impairment in motor function in preterm-born infants.

1   Introduction

Every year, an estimated 2.2 million babies worldwide are born very preterm
(born at 32 weeks gestation or younger) [6]. Very preterm birth puts newborns
at a high risk of long-term motor dysfunction (e.g. cerebral palsy), which places
significant burdens on the child, the family and the community [2, 12]. Early
detection of motor dysfunction could enable more rapid identification of infants
who would benefit from rehabilitative interventions. While motor outcomes can
be assessed in preterm-born infants at 18 months of age using the Bayley Scales of
Infant and Toddler Development, Third Edition (Bayley-III) [4], we desire earlier
identification of infants at risk in order to inform care and ongoing monitoring.

Certain brain pathologies, such as white matter injury (WMI) and intra-
ventricular hemorrhaging (IVH) are detectable in a structural MRI scan of an
infant's brain. It is also known that some of these pathologies, as well as findings
from more advanced MR methods such as diffusion tensor imaging (DTI), are
associated with later neurodevelopmental outcomes [3, 8]. However, most studies

to date have focused on group differences or correlations between specific DTI measures and motor outcomes. For example, Chau et al. recently reported that the trajectory of brain maturation from early in life to term-equivalent age, using region-of-interest-based DTI measures, was associated with neurodevelopmental outcomes [8]. Ball et al. examined the relationship between connectivity in the thalamo-cortical connectome in preterm infants and Bayley-III scores [3]. They found that the strength of certain connections were significantly correlated with outcomes.

In comparison to finding correlations, prediction is a harder problem. In order to predict accurately, the complete set of factors contributing to outcome must be modelled. Prediction of motor function from brain structure at birth is thus very challenging due to the large number of confounding factors affecting brain development, including potentially unknown genetic and environmental factors. It is especially difficult in young infants due to the combination of limited image resolution and small brain sizes, artifacts due to motion and rapid structural change across a small temporal window [5]. Furthermore, datasets are often class-imbalanced, containing fewer abnormal cases. This last issue is of particular importance to prediction since many prediction models are highly sensitive to imbalanced training data [10].

Strategies exist in the literature to alleviate this class imbalance problem by augmenting the training set. For example, the synthetic minority over-sampling technique (SMOTE) finds K nearest neighbours to each training instance and interpolates new instances randomly along lines connecting neighbours [9]. Another method is to sample synthetic instances from an approximate distribution of positive instances, learned using kernel density estimation (KDE). Alternatively, the dominant modes of variation for positive training instances can be learned using principal component analysis (PCA), then sampled to generate new instances. The method that generates the most realistic synthetic instances (i.e., those which improve prediction accuracy the most) is likely application dependent, making class imbalance challenging for prediction problems.

Despite the challenges, prediction of long-term motor dysfunction within the first few weeks of birth remains a desirable goal as it would enable better treatment planning and a more informed assessment of patient outcome. Recently, Ziv et al. used connectome based features from term infants scanned at 6 months of age to predict general neurological outcome at 12 months [18]. Here we set the goal of predicting motor outcome at 18 months from scans taken within the first weeks after birth. In our study, the earlier post-menstrual age (PMA) at scan and the larger temporal gap between scan and outcome makes our task even more challenging. Similar to Ziv et al. we use dimensionality-reduced connectome features and a support vector machine (SVM) classifier to achieve this goal. However, we also introduce a novel method for generating synthetic training samples designed to tackle the specific challenges in our data. Our training data contains only a small number of positive instances existing in a high-dimensional space, likely constrained to a complex manifold. In order to only generate re-
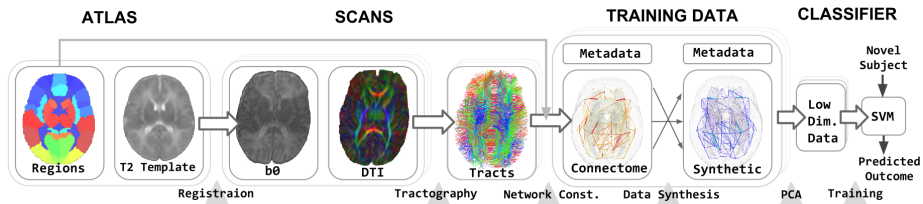
Fig. 1. High level schematic representation of connectome and training pipeline.

alistic data, we propose a data interpolation technique that generates synthetic instances which are restricted to be more similar to individual known instances.

In this paper we show for the first time that 18-month motor outcomes, assessed with the Bayley-III, can be predicted from an MRI taken in the first weeks of a preterm neonate's life. Our method achieves an accuracy of $> 70\%$, which establishes a baseline level of accuracy for this important yet very challenging task. Further, we demonstrate that data augmentation can reliably improve prediction scores and that our novel method for generating local synthetic instances outperforms competing methods.

2   Methods

In Fig. 1, we present a schematic diagram of our supervised learning framework for predicting motor function at 18 months from DTI scans acquired in the first weeks of life. Details for this pipeline are presented below.

Dataset: The cohort used in this study was a group of 115 preterm infants born between 24 and 32 weeks PMA. Neonates underwent a brain MRI between 27 and 45 weeks PMA on a Siemens (Erlangen, Germany) 1.5T Avanto using VB 13A software. Each scan was a multi-slice 2D axial EPI diffusion MR acquisition (TR 4900 ms; TE 104 ms; FOV 160 mm; slice thickness, 3 mm; no gap) with 3 averages of 12 non-colinear gradient directions, with an isotropic in-plane resolution of 0.625 mm. Two such acquisitions were acquired: one at $b = 600$ s/mm^2 and one at $b = 700$ s/mm^2. The combined diffusion weighted image set was preprocessed using the FSL Diffusion Toolbox (FDT) pipeline[1] and tensors were fit using RESTORE [7]. Nearly half of the subjects (53) were scanned twice for a total of 168 diffusion tensor images.

At 18 months, each subject was evaluated using the Bayley-III test which produces three composite scores of cognitive, language and motor skills [4]. The scores are normalized with mean of 100 and standard deviation of 15; we considered adverse motor outcomes as scores below 85 (i.e., -1 std.). In our cohort, 146 scans were of infants with normal motor function at 18 months, and 22 scans were of infants with abnormal motor function.

---

[1] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FDT

Connectome Construction and Analysis: Each DTI was segmented using a neonatal atlas of 90 brain regions from the IDEA group at University of North Carolina (UNC) School of Medicine, Chapel Hill [15]. The atlas' associated T2 template was aligned to the b0 image of each DTI scan in order to segment each brain. Alignment was performed as a rigid registration using FMRIB's Linear Image Registration Tool (FLIRT) [11] followed by a deformable registration using the MATLAB Image Registration Toolbox (MIRT)[2]. Full-brain tractography was performed on each scan using TrackVis [16]. Each connectome was constructed by grouping tracts according to their endpoint regions. Following [5], we constructed three connectome types for each scan, including a mean-FA weighted connectome, a tract-count connectome and a normalized tract-count connectome. Network measures summarizing the connectome's topological properties were computed for each connectome type. Ten individual network measures were extracted: 1) mean nodal degree, 2) transitivity, 3) global efficiency, 4,5) raw and normalized modularity, 6,7) raw and normalized clustering coefficient, 8,9) raw and normalized characteristic path length and 10) small-worldness. Each measure was computed for mean-FA, tract-count and normalized tract-count connectomes giving a total of $30 = 10 \times 3$ measures. For a comprehensive summary of network measures and their meanings, see [13].

Classification: For each DTI scan, metadata and connectome features were extracted. See Table 1 for a full list of feature types. Metadata features included gender, age-at-scan and age-at-birth. Age-at-scan is included because we expect it to be an important co-variate given the rapid development of the brain across the age range of our cohort. Connectome features included mean FA across each of the $4005 = 90 \times 89/2$ edges and other high-level network features as described above. WMI severity (graded $[0,3]$) and IVH severity (graded $[0,4]$), assessed from a T1 MRI by an experienced neuroradiologist (KJP), were included as additional features. All features were concatenated into one feature vector.

Feature vectors from the training set were processed using PCA to extract, at most, the top m modes of variation. At test time, instances were projected into this PCA space. A linear dimensionality reduction method was used instead of, for instance, a kernel based method in order to prevent over-fitting to our sparsely sampled, high-dimensional training set. An SVM classifier was then trained on the instances in this reduced space.

Local Synthetic Instances (LSI): Due to the limited number of cases in our cohort that show motor dysfunction (22 out of 168), our training set has a severe class imbalance ($> 6{:}1$). We compensate for this by both replicating existing positive instances, $\mathbf{t}_i$ (i.e., feature vectors), and by generating new synthetic positive instances, $\mathbf{s}_j$. New instances are generated by interpolating instance feature vectors within the training set of N positive instances.

We do not know the true distribution of positive instances but we assume that it is locally smooth. Under this assumption we can generate reasonable

---

[2] https://sites.google.com/site/myronenko/research/mirt

synthetic instances as long as they are near existing positive training instances. To this end, we seek an interpolant that satisfies two conditions: (i) gives the majority of weight to one training instance and (ii) gives some non-zero weight to other instances to ensure some variation. We achieve this by randomly assigning elements from a normalized p-series as weights to each instance. Let P^j be a random $N \times N$ permutation matrix and $\mathbf{r}^j = P^j[1, 2, ..., N]^T$. Then, the elements, r_i^j, of $\mathbf{r}^j$ are numbers from 1 to N and the vector $\mathbf{r}^j$ is randomly ordered and,

$$\mathbf{s}_j = \sum_{i=1}^{N} w(r_i^j)\mathbf{t}_i \qquad \text{where} \qquad w(i) = \frac{i^{-p}}{\sum_{n=1}^{N} n^{-p}}. \qquad (1)$$

Thus, for the jth synthetic instance, $\mathbf{s}_j$, $\mathbf{r}^j$ randomly assigns weights to training instances without replacement. Note that $\sum_{i=1}^{N} w(i) = 1$. Also, note that $\forall i > 1, w(1) \geq 2^p w(i)$ and for $p = 2$ (and in fact, any $p > 1.7287$), $w(1) > \sum_{i=2}^{N} w(i)$ for any N. In other words, the weight on one instance dominates the others and each synthetic instance will be generated local to one training instance as desired. Fig. 2a is a schematic example of synthetic instances generated from three real instances.

The proposed LSI method has desirable properties compared to other existing methods in the case of high-dimensional data and small number of training instances, as we have here. For instance, while samples from a KDE-estimated distribution are likely to be near training instances, a fixed kernel is used, resulting in new instances that may not vary away from training instances in a realistic way. Instead, LSI encourages new instances to be near training instances but requires variation to be towards known training data. Furthermore, unlike KDE, LSI does not allow extrapolation. This is adventageous since, given a small training size, extrapolation is unlikely to yield realistic samples (Fig. 2b(i)). SMOTE generates instances local to a subset of known instances but fails where the manifold of positive instances is too sparsely sampled to be approximated using nearest neighbours (Fig. 2b(ii)). Sampling from PCA modes ensures that new instances only vary along primary modes of variation but may generate samples that are anatomically implausible since they may not be near any training instance (Fig. 2b(iii)). In comparison, the proposed method offers a balance between trusting local instances and using global information.

3   Results

We evaluated our method using a variety of feature subsets and instance synthesis methods. We assessed classification accuracy for different sets of feature types (Table 1) via 1000 rounds of cross-validation. In each round, one positive and one negative instance were left out for testing. Test subjects were omitted from training data and, since some subjects were scanned twice, the total training set varied between 164 and 166 scans. For these tests, we set $m = 20$, which explains $> 99\%$ of the variance and set $p = 2$. SVM misclassification penalty was set empirically to $C = 2^{-7}$ but test accuracy was relatively insensitive to this value, varying only about 1% for $2^{-8} < C < 2^{-6}$. Classes

were balanced by weighting positive instances via replication. Table 2 shows prediction results for different sets of features and compares those tests run with synthetic training data versus those without. In the tests with synthetic data, the training set was first doubled by generating synthetic instances, then real positive training instances were replicated until the classes were balanced.
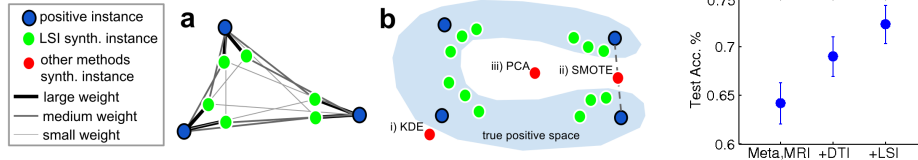

Fig. 2. a) Schematic representation of LSI weights for 6 synthesized instances. b) Possible failure cases for other data augmentation methods.

Fig. 3. Test accuracy and 95% CIs for select tests from Table 2.

Table 1. Name and description of each feature type.

Name	Description (# features)	Name	Description (# features)
Meta	Birth age, scan age and gender (3)	MRI	WMI and IVH scores at birth (2)
Edge	Connectome edge FA values (4005)	Network	High-level network measures (30)

Table 2. Mean training (Tr) and test accuracy, sensitivity (Sn) and specificity (Sp) for 1000 rounds of leave-2-out cross validation. Tests marked * are plotted in Fig 3. Best test accuracy is in bold.

Feature Types Used	No Synth. Data				With Synth. Data			
	Tr	Test	Sn	Sp	Tr	Test	Sn	Sp
Meta, MRI	69.0	64.2*	0.46	0.83	69.9	64.4	0.47	0.82
Meta, Edge	85.0	58.5	0.33	0.84	86.1	58.5	0.34	0.83
Meta, Edge, Network	85.5	61.2	0.34	0.82	85.7	62.8	0.43	0.83
Meta, Edge, Network, MRI	84.4	64.7	0.51	0.79	85.0	64.7	0.49	0.81
Meta, Network, MRI	77.5	69.0*	0.56	0.82	79.3	**72.3***	0.66	0.79

Note that the addition of synthetically-generated instances improved our highest test accuracy by 3%. Furthermore, as expected, the inclusion of high-level connectome features and MRI-based gradings (i.e. WMI & IVH grades) consistently boosted accuracy on average by about 3% each. Interestingly, *excluding* edge FA features causes the classification accuracy to *improve* when network measures and MRI-based scores are included. This is likely because the edge features are noisy and that much of the relevant structural information from the edge features is captured more succinctly in the network measures, WMI grade, and IVH grade. It may also suggest that no single white-matter fibre bundle is strongly tied to motor-outcome and instead that more widespread factors are at work. This finding is consistent with recent work in [2] which showed that the causes of poor motor outcome are multi-faceted. Note also that while metadata and MRI based information alone provides reasonable predictive power, it is clearly advantageous to include connectome information derived

from DTI. This is shown in Fig. 3 with 95% confidence intervals (CI) for each result.

For the test using the subset of features that gave the highest test accuracy (Meta, Network, MRI), test accuracy was 69% across an infant's first scans and 74% across their second scans. This suggests that the later scans may be at some advantage in their ability to predict outcome. This is not surprising since the temporal gap between scan-time and assessment for these scans is smaller.

Again using the same features and ratio of synthetic to training instances that achieved the highest accuracy in the above experiment, we tested four different instance synthesis methods (Table 3). For KDE, a Gaussian kernel was used and, after checking a range of scales, a standard deviation equal to the mean distance between samples was found to give the highest accuracy. SMOTE with different costs (SDC) is a variant of SMOTE, improved to better handle imbalanced classes [1]. For SDC, we set $K = 1$ as it was found to give highest accuracy over range $[1, 5]$. For PCA, once the variation modes were learned, new instances were generated by sampling the Gaussian distribution defined by the variation modes. To be consistent with the PCA step used for classification, again the top $m = 20$ modes of variation were used. We also compared these synthesis methods against subset sampling optimization (SSO) [17], a state-of-the-art undersampling method, and weighted Lagrangian twin SVM (WLT), a classifier designed to natively deal with class imbalance [14]. No replicated or synthetic instances were used with these methods. Note that since we use non-image data in our feature vector, data augmentation methods which modify the images directly are not applicable here, and so were not tested. Table 3 shows that our proposed method outperforms the competing methods. These findings agree with our hypothesis at the end of Section 2, that LSI is well suited to a small sample size from a complex manifold in high-dimensional space.

Table 3. Comparison between class balancing methods. Test accuracies with 95% CIs plotted on right.

Method	Tr	Test	Sn	Sp	Method	Tr	Test	Sn	Sp
WLT	74.2	54.3	0.23	0.85	PCA	79.0	68.5	0.57	0.80
SSO	88.8	62.1	0.56	0.68	KDE	79.7	68.9	0.57	0.81
SDC	78.3	67.6	0.56	0.80	LSI	79.3	**72.3**	0.66	0.79

4   Conclusions

In this paper we predicted preterm infant motor outcomes at 18 months using structural connectome features from DTI scans taken at birth. In doing so, we established a baseline accuracy of over 70% for this challenging but important task. We also proposed a novel method to mitigate the effects of small positive sample sizes common to normal/abnormal datasets. Our approach improved prediction accuracy and outperformed a variety of other methods for this application. In future works, we plan to more thoroughly explore the characteristics of our LSI method and find ways (e.g. via other engineered or learned features and different machine learning techniques) of improving prediction accuracy even further.

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Machine Learning: ECML 2004, pp. 39–50. Springer (2004)
2. Back, S.A., Miller, S.P.: Brain injury in premature neonates: A primary cerebral dysmaturation disorder? Annals of neurology 75(4), 469–486 (2014)
3. Ball, G., Pazderova, L., Chew, A., Tusor, N., Merchant, N., Arichi, T., Allsop, J.M., Cowan, F.M., Edwards, A.D., Counsell, S.J.: Thalamocortical connectivity predicts cognition in children born preterm. Cerebral Cortex p. bhu331 (2015)
4. Bayley, N.: Manual for the Bayley Scales of Infant Development. Harcourt, San Antonio, 3rd edn. (2006)
5. Brown, C.J., Miller, S.P., Booth, B.G., Andrews, S., Chau, V., Poskitt, K.J., Hamarneh, G.: Structural network analysis of brain development in young preterm neonates. NeuroImage 101, 667–680 (2014)
6. C P Howson, M V Kinney, J.E.L.: Born too soon: The global action report on preterm birth. World health organization. Geneva (2012)
7. Chang, L.C., Jones, D.K., Pierpaoli, C.: RESTORE: Robust estimation of tensors by outlier rejection. Magnetic Resonance in Medicine 53, 1088–1095 (2005)
8. Chau, V., Synnes, A., Grunau, R.E., Poskitt, K.J., Brant, R., Miller, S.P.: Abnormal brain maturation in preterm neonates associated with adverse developmental outcomes. Neurology 81(24), 2082–2089 (2013)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of AI research 16(1), 321–357 (2002)
10. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429–449 (2002)
11. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17(2), 825–841 (2002)
12. Miller, S.P., Ferriero, D.M., Leonard, C., Piecuch, R., Glidden, D.V., Partridge, J.C., Perez, M., Mukherjee, P., Vigneron, D.B., Barkovich, A.J.: Early brain injury in premature newborns detected with mri is associated with adverse early neurodevelopmental outcome. The Journal of pediatrics 147(5), 609–616 (2005)
13. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52(3), 1059 – 1069 (2010)
14. Shao, Y.H., Chen, W.J., Zhang, J.J., Wang, Z., Deng, N.Y.: An efficient weighted lagrangian twin support vector machine for imbalanced data classification. Pattern Recognition 47(9), 3158–3167 (2014)
15. Shi, F., Yap, P.T., Wu, G., Jia, H., Gilmore, J.H., Lin, W., Shen, D.: Infant brain atlases from neonates to 1-and 2-year-olds. PLoS One 6(4), e18746 (2011)
16. Wang, R., Benner, T., Sorensen, A.G., Wedeen, V.J.: Diffusion toolkit: a software package for diffusion imaging data processing and tractography. In: Proc Intl Soc Mag Reson Med. vol. 15, p. 3720 (2007)
17. Yang, P., Zhang, Z., Zhou, B.B., Zomaya, A.Y.: Sample subset optimization for classifying imbalanced biological data. In: Advances in Knowledge Discovery and Data Mining, pp. 333–344. Springer (2011)
18. Ziv, E., Tymofiyeva, O., Ferriero, D.M., Barkovich, A.J., Hess, C.P., Xu, D.: A machine learning approach to automated structural network analysis: application to neonatal encephalopathy. PloS one 8(11), e78824 (2013)