


RESEARCH ARTICLE

Open Access



# Autism risk classification using placental chorionic surface vascular network features

Jen-Mei Chang<sup>1\*</sup> , Hui Zeng<sup>1</sup>, Ruxu Han<sup>1</sup>, Ya-Mei Chang<sup>2</sup>, Ruchit Shah<sup>3</sup>, Carolyn M. Salafia<sup>3,4,5</sup>, Craig Newschaffer<sup>6</sup>, Richard K. Miller<sup>5,7</sup>, Philip Katzman<sup>5,7</sup>, Jack Moye<sup>8</sup>, Margaret Fallin<sup>9</sup>, Cheryl K. Walker<sup>5,10</sup> and Lisa Croen<sup>5,11</sup>

## Abstract

**Background:** Autism Spectrum Disorder (ASD) is one of the fastest-growing developmental disorders in the United States. It was hypothesized that variations in the placental chorionic surface vascular network (PCSVN) structure may reflect both the overall effects of genetic and environmentally regulated variations in branching morphogenesis within the conceptus and the fetus' vital organs. This paper provides sound evidences to support the study of ASD risks with PCSVN through a combination of feature-selection and classification algorithms.

**Methods:** Twenty eight arterial and 8 shape-based PCSVN attributes from a high-risk ASD cohort of 89 placentas and a population-based cohort of 201 placentas were examined for ranked relevance using a modified version of the random forest algorithm, called the Boruta method. Principal component analysis (PCA) was applied to isolate principal effects of arterial growth on the fetal surface of the placenta. Linear discriminant analysis (LDA) with a 10-fold cross validation was performed to establish error statistics.

**Results:** The Boruta method selected 15 arterial attributes as relevant, implying the difference in high and low ASD risk can be explained by the arterial features alone. The five principal features obtained through PCA, which accounted for about 88% of the data variability, indicated that PCSVNs associated with placentas of high-risk ASD pregnancies generally had fewer branch points, thicker and less tortuous arteries, better extension to the surface boundary, and smaller branch angles than their population-based counterparts.

**Conclusion:** We developed a set of methods to explain major PCSVN differences between placentas associated with high risk ASD pregnancies and those selected from the general population. The research paradigm presented can be generalized to study connections between PCSVN features and other maternal and fetal outcomes such as gestational diabetes and hypertension.

**Keywords:** Placental chorionic surface vascular network (PCSVN), Autism spectrum disorder risk, Boruta algorithm, Linear discriminant analysis, Placenta, Principal component analysis, Random forest, Arterial network

## Background

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with deficits in three defining areas: social reciprocity, communication, and restricted and repetitive patterns of behaviors. Symptoms are typically developed by 36 months of age. The causes of ASD are not definitive and include both genetic and non-inherited factors and exposures. About one in 68 children in the United

States and one percent of the world population has been identified with ASD, according to a 2016 estimate from the Center for Disease Control [1]. The lifelong cost of ASD in the United States is about \$2.4 million for a person with an intellectual disability, or \$1.4 million for a person without intellectual disability [2]. Since the brain is most responsive to treatment in the first year of life, early intervention is key to help children diagnosed with ASD. However, since most of the diagnoses of ASD are not made until the child is three or four years old, the best opportunities for intervention have already been lost. There is no doubt that ASD is a global epidemic and efforts are needed

\*Correspondence: jen-mei.chang@csulb.edu

<sup>1</sup>Department of Mathematics and Statistics, California State University, Long Beach, CA 90840-1001 Long Beach, USA

Full list of author information is available at the end of the article

in developing reliable bio-markers in assessing prenatal and neonatal risk to not only increase the effectiveness of the treatments and minimize the cost to treat children with ASD.

One way to develop a bio-marker is to study groups of children exposed to high risk for ASD. For example, children with a twin sibling have a much higher chance of getting diagnosed with ASD. In particular, studies have shown that among identical twins, if one child has ASD, then the other will be affected about 36–95% of the time. In non-identical twins, if one child has ASD, then the other is affected about 0–31% of the time [3, 4]. Moreover, parents who have a child with ASD have a 2–18% chance of having a second child who is also affected [5, 6]. Based on a research study completed by the Baby Siblings Research Consortium [7], the recurrence risk of ASD was 18.7% for families with at least one older sibling with ASD. Children with more than one older sibling with ASD were even more likely to be diagnosed, with a 32.2% risk – twice that of children with only one older autistic sibling [7].

As we know that the gene families that control branching morphogenesis in the permanent organs such as kidneys, lungs, and pancreas are related to the genes that control branching morphogenesis in placenta [8], this makes placenta an ideal organ to study fetal vasculogenesis and angiogenesis. Abnormal placental angiogenesis and vasculogenesis underly a number of pregnancy complications, from preeclampsia to fetal growth restriction and pre-term birth [9–11]. Evidence suggests that it may also be responsible for irregular placental shape [12, 13]. A major feature of the whole placenta, the placental chorionic surface vascular network (PCSVN), has not been extensively studied due to the extreme difficulty in reliably extracting PCSVN features from digital images of the fetal surface [14]. It was hypothesized that variation in PCSVN structure, the template of the fetal organ positioned at the interface of the mother and the conceptus, may reflect both the overall effects of genetic and/or environmentally regulated (e.g., [15]) variations in branching morphogenesis within the conceptus, and may also mirror vascular network alterations in the fetus' vital organs.

Although there were results linking chorionic surface shapes to immediate neonatal outcomes such as birth weight after adjustment for gestational length [13, 16], very limited work has been done on the connection between PCSVN features and neonatal outcomes. Preliminary research results [17] suggested that there are significant differences in PCSVN features (e.g., Number of branch generations and angles of vascular branching.) in children at increased risk for autism spectrum disorder. The study was conducted on a data set of 109 placentas with 33 from a high-risk ASD cohort and 76 from a population-based cohort, and did not include

a mechanism to classify a given placenta as high-risk for ASD with the PCSVN features that were deemed significant.

Our goal in this paper is to provide sound evidences to support the study of ASD risks through the placental chorionic surface vascular networks. We will do so by developing a set of methods to explain major PCSVN differences between placentas associated with high risk ASD pregnancies and those selected from the general population. The methods assume no a-priori knowledge on which factors might have been in play to establish the difference and can be generalized naturally to other maternal and fetal outcomes such as gestational diabetes and hypertension. A flowchart of our proposed work is given in Fig. 1. We will begin by describing the ways we obtain PCSVN features from a digital photograph of a placenta (preprocessing stage of Fig. 1), then discussing the methods we use to distill a relatively large set of PCSVN features into a subset of physically meaningful ones (feature selection stage of Fig. 1), and finally presenting the way the ASD risk is assigned (classification stage of Fig. 1).

## Methods

### Data sets

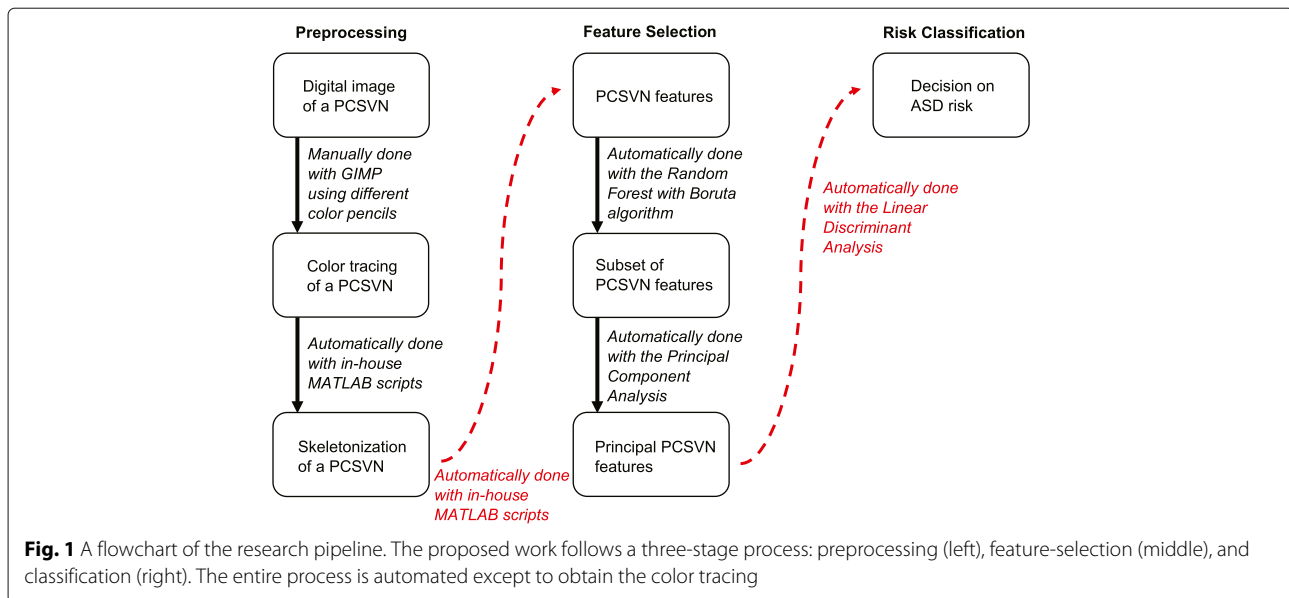
The placentas investigated in this study are taken from two independently collected cohorts, Early Autism Risk Longitudinal Investigation (EARLI) [18] and National Children's Study (NCS). Protocols for the original data collections were approved by the pertinent Institutional Review Boards. This study concerns with secondary analysis on de-identified data.

EARLI is an autism enriched-risk pregnancy cohort that focuses on the prenatal and early life periods of children who have biological siblings already diagnosed with ASD. EARLI children are at an increased risk for ASD. On the other hand, NCS is a population-based cohort with pregnancies at unknown risk for ASD. NCS was designed to study environmental influences on child health and development and it enlisted participants without a bias towards risks and diagnoses in autism. Placentas in NCS are used here as an unselected low-risk baseline. We randomly selected 201 placentas from NCS and 89 placentas from EARLI in this study.

We have limited clinical data such as gender, gestation age, placental weight, and birth weights on small subsets of NCS and EARLI. Our data sets will be reduced significantly if we were to include these clinical attributes in the study; hence, the present study concerns only with the connections between vascular features of the PCSVN and risk outcomes for ASD.

### Vascular features

Digital photographs of the fetal surface were obtained on 201 NCS placentas and 89 EARLI placentas following the



same imaging protocol (e.g., Fig. 2a). The photos of the placentas were taken either at delivery or upon pathology evaluation with fresh tissue. The raw PCSVN images in both NCS and EARLI data sets were captured using the same camera and polarizing filter. The distance between the camera and the placenta being imaged was fixed in NCS while there was a slight variability among EARLI images. Lighting condition was also fixed in NCS while there was a slight variability in lighting among EARLI images.

PCSVN of each placenta was first traced manually (e.g., Fig. 2b) following the protocol documented in [14] using GIMP by one of the researchers who was blind to the risk categories. To make the manual tracing consistent and compatible with the computer algorithms, the researchers in [14] developed a protocol in which different colors and pencil sizes were used to mark different vessel thicknesses and separate the placental chorionic surface arterial network from the subjacent venous network. All tracings were reviewed for consistency and checked by a single researcher. Ten percent of the tracings were selected at random and traced by a second tracer, to confirm and maintain our high inter-rater reliability. Since our study relied on the color tracings of the PCSVN instead of the original raw images, the slight variation in the image acquisition process should pose little concern to the validity of our results as long as the images were clear enough for the tracer to identify the location of the vessels.

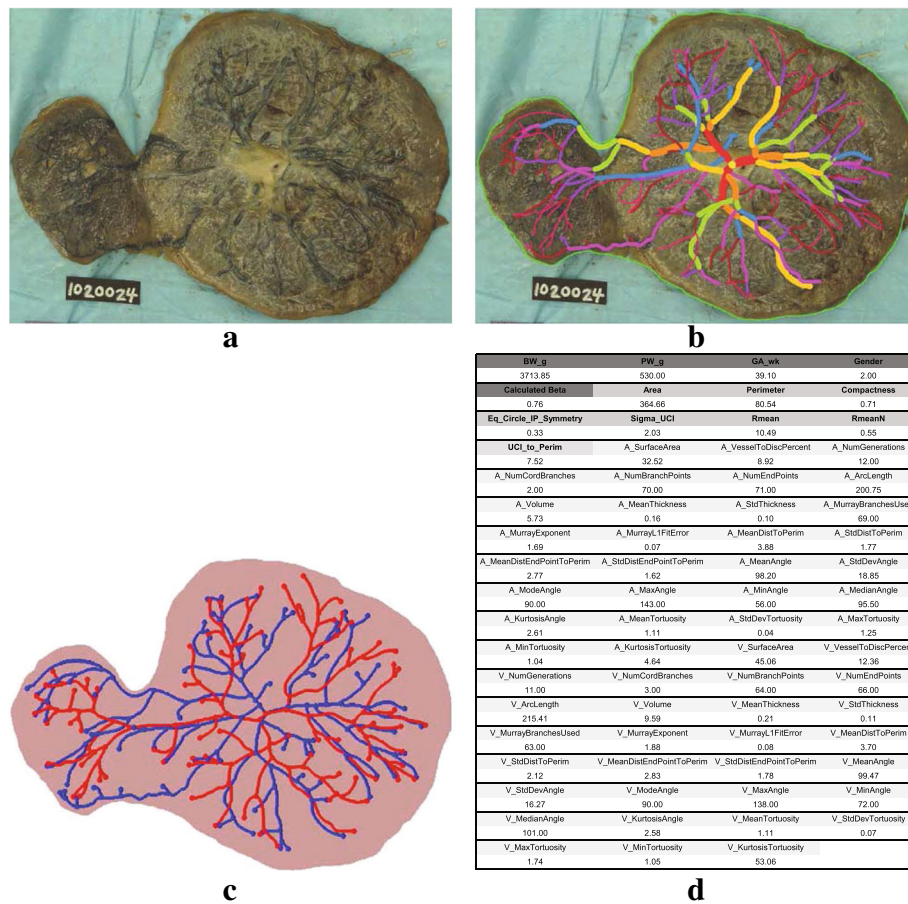
Color tracings were uniformly scaled and converted to  $1380 \times 1440$  pixel binary images so the width of the vessels were normalized. One centimeter was marked with two blue dots on the ruler within the original photograph of the placenta to give scale. Roughly 35 pixels in the digital image corresponded to 1 cm on the placenta. Tracings

were aligned so that the umbilical cord insertion lies at the center of the image. Each traced image was then fed through a series of MATLAB scripts, written completely by the researchers, to produce a fully connected graph network (e.g., Fig. 2c) based on its color profile. Notice that, for example, in Fig. 2c each branch point is marked with a solid dot to help with any calculation related to branch points. The 1-pixel-wide skeleton graphs were then used to produce 64 numerical values (e.g., Fig. 2d), of which eight are shape-related (e.g., perimeter and area of placental chorionic surface plate) and 56 are vessel-related (e.g., number of branch points and vessel length).

Within the 56 vascular features, half of them were calculated on arterial networks and the other half were done on venous networks. Those features can be generally classified as counting descriptors (e.g., number of branches), measuring descriptors (e.g., arterial length), and relating descriptors (e.g., the distance between vessel and plate boundary). While similar analyses and results are available on the venous network, we will only present results on the arterial network here. The arterial networks are typically much more identifiable and visible than the venous networks. This allows the tracer to trace the arterial networks with a much higher level of precision and accuracy [14].

#### Boruta algorithm for relevant feature selection

The Boruta algorithm is a feature selection algorithm for finding a minimal set of relevant variables. This method, which builds around the concept of random forest and decision trees [19], systematically and iteratively removes features that are less relevant than random probes by a statistical test [20–22]. By adding randomness to the system and collecting results from the ensemble of randomized samples one can reduce the misleading impact of random



**Fig. 2** The process of obtaining a feature vector for each placenta. **a** A digital photograph of the placental chorionic surface vascular network (PCSVN) from the NCS data set. **b** Traced PCSVN for the image in (a) following the tracing protocols in [14]. **c** The skeletonisation of the traced PCSVN image in (b) that was produced by a MATLAB program written in house by the research team. **d** Numerical values of PCSVN features computed by our MATLAB program for the image in (c). Each of the 290 placentas in our data set is associated with a list of values similar to those given in (d)

fluctuations and correlations and reduce the undesirable effect of over-fitting.

In the Boruta algorithm, each attribute has a “shadow attribute” which is created by shuffling the values of the original attribute. During a single run of Boruta, a feature attribute is deemed important if its importance score (z-score) is significantly bigger than the maximum z-score among all shadow attributes (MZSA). It is deemed unimportant if its z-score is significantly lower than the MZSA. A two-sided test of equality with the MZSA is performed on feature attributes that have undetermined importance. This process is repeated until the importance is assigned for all attributes or the algorithm has reached the previously set number of random forest runs, which is 500 in our simulations.

In conclusion of a Boruta simulation, a ranked list of features, ordered by their importance measure given in z-scores is produced. A major advantage of the Boruta

strategy is its ability to discern truly important features from those that gain importance due to random correlations in data. Consequently, it gives us a powerful tool to establish a hierarchy of relevance when we need to study biological factors of various nature. In the current study, the Boruta algorithm allows us to confidently identify a list of PCSVN features that characterize the difference between high- and low-risk ASD placentas.

### Reduce dimensionality and collinearity with principal component analysis

Many features extracted from the Boruta algorithm remain correlated, making it harder to interpret the fundamental principles that govern villous growth. To this end, we used Principal Component Analysis (PCA) to reduce that collinearity, bringing a moderately large number of features down to a few independent signatures. These linearly independent components will be ranked

by their proportion of contribution to the data variance. Precisely, if we let

$$F = \begin{bmatrix} | & | & & | \\ \mathbf{f}_1 & \mathbf{f}_2 & \cdots & \mathbf{f}_p \\ | & | & & | \end{bmatrix}$$

be the  $N \times p$  feature matrix, where  $p$  = number of samples (290 in this case) and  $N$  = the number of significant features selected by the Boruta algorithm, then the  $N \times N$  covariance matrix  $C = \frac{1}{N-1} \tilde{F} \tilde{F}^T$  gives the feature variance on the diagonal entries and co-variance on the off-diagonal entries, where  $\tilde{F}$  is centered around the feature mean. When  $\tilde{F}$  is factored through its reduced singular value decomposition (SVD),  $\tilde{F} = USV^T$ , where  $U = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k]$  is  $N \times k$ ,  $S$  is  $k \times k$ , and  $V$  is  $p \times k$ , the best feature basis (hence, the best feature space) to represent the data in the reduced  $k$ -dimensional space is stored in the  $k$  column vectors of  $U$  with  $k \ll N$ . The best choice of  $k$  depends on how much variance we wish to capture. By representing the original data points through this new set of coordinates  $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ , the reduced-dimension data points,  $D = U^T \tilde{F}$ , are now expressed by a set of linearly independent principal components.

This allows us to investigate physical interpretations of these  $N$  features by finding which variables correlate most strongly with each component, i.e., finding which numbers are large in magnitude or the farthest away from zero in either positive or negative direction. Variables of large magnitude within the same principal component vary together, i.e., if one increases, then the remaining ones also increase. Thus, PCA was used to identify groups of biological effects of villous growth as a consequence of ASD risk.

### Classification with linear discriminant analysis

Associate each placenta in the data set with a  $k$ -dimensional vector,  $\mathbf{p}$ , where each entry of  $\mathbf{p}$  corresponds to a principal component coordinate. That is,  $\mathbf{p}$  is a column vector in the matrix  $D = U^T \tilde{F}$  in the previous section. To classify high-risk ASD placentas, Linear Discriminant Analysis (LDA) was conducted on the set of 290 placentas represented in the PCA coordinates. Suppose  $D_1$  and  $D_2$  are sets of PCA-reduced data points of low-risk and high-risk ASD placentas, respectively. Linear discriminant analysis amounts to finding a projection direction  $\mathbf{w}_{\text{opt}}$  that maximizes the *between-class* scatter and minimizes the *within-class* scatter among the data points, which is equivalent to solving the optimization problem,

$$\mathbf{w}_{\text{opt}} = \operatorname{argmax}_{\|\mathbf{w}\|=1} \frac{(\mathbf{w}^T \mathbf{m}_2 - \mathbf{w}^T \mathbf{m}_1)^2}{S_1^2 + S_2^2},$$

where  $\mathbf{m}_i$  is the  $i$ th class mean and  $S_i^2 = \sum_{\mathbf{y} \in D_i} (\mathbf{w}^T \mathbf{y} - \mathbf{w}^T \mathbf{m}_i)^2$  is the within-class scatter among the  $i$ th class.

The optimization problem is then solved through its matrix form:  $\mathbf{w}_{\text{opt}} = \operatorname{argmax} J(\mathbf{w})$ , where  $J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} = \frac{N(\mathbf{w})}{D(\mathbf{w})}$ . The between-class scatter matrix is given by  $S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$  and the within-class scatter matrix is given by  $S_W = \sum_{i=1,2} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$ .  $\mathbf{w}_{\text{opt}}$  is the largest eigenvector associated with the largest eigenvalue to the generalized eigen-problem  $S_B \mathbf{w} = \lambda S_W \mathbf{w}$ . This problem can be solved numerically through an SVD-based method.

Assume the projected values of the points in  $D_1$  fall to the left of those in  $D_2$ . If we set the separation threshold,  $\alpha$ , to be  $\frac{1}{2} (\min \{\mathbf{w}_{\text{opt}}^T D_2\} + \max \{\mathbf{w}_{\text{opt}}^T D_1\})$ , then a given placenta,  $\mathbf{p}$ , is labeled low-risk for ASD if  $\mathbf{w}_{\text{opt}}^T \mathbf{p} \leq \alpha$  and labeled high-risk for ASD if  $\mathbf{w}_{\text{opt}}^T \mathbf{p} > \alpha$ .

To generate error statistics, we perform LDA with a 10-fold cross validation. Essentially, the entire data set was randomly split into ten disjoint groups where each group of 29 placentas was used as testing probes to produce error statistics while the rest of the data set was used to find  $\mathbf{w}_{\text{opt}}$  during each trial. Sample population (i.e., 30.69% of the population are high-risk for ASD and 69.31% are low-risk for ASD) was used in the model as an estimated priors to confirm that the use of Linear (instead of quadratic) Discriminant Analysis was the correct model.

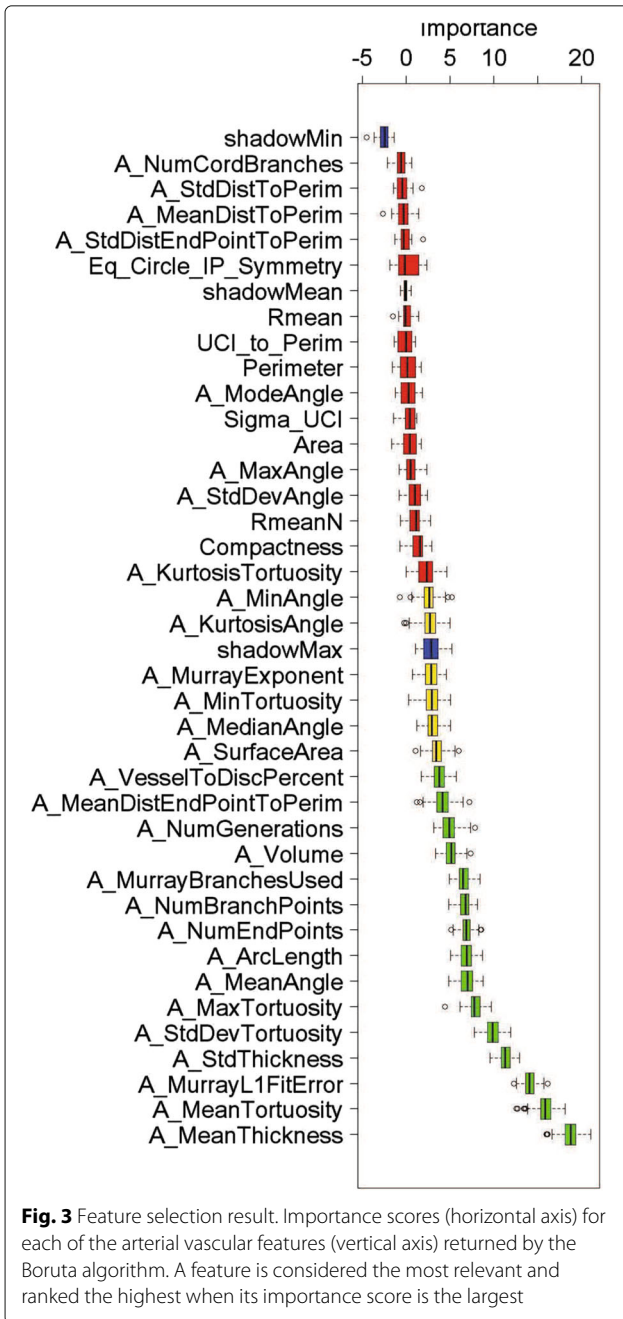
## Results

### Feature selection and dimensionality reduction

The Boruta algorithm selected 15 arterial features. Figure 3 gives a visual output from running the Boruta algorithm in the programming language R. The box plot of each attribute, listed from the lowest (top) to the highest (bottom) rank, was the result of the  $z$ -score spread obtained from running the random forest algorithm 500 times. The 15 relevant and important features selected by Boruta are given in the ‘‘Vascular features’’ column of Table 1 and appear in green in Fig. 3.

PCA, implemented in MATLAB, was applied to the Boruta-selected 15-feature set. Five Principal Components (PCs) were retained to capture roughly 88% of the data variability. The principal components, also known as the eigenvectors of the covariance matrix, are given in Table 1 to delineate the source of contribution for each principal direction. Notice that many attributes within the same principal component are correlated. Next, we examine closely on the mathematical relationships of these features and deduce a list of independent principal features that will explain the biological and structural difference between the two cohorts. Here, we chose to borrow the term ‘‘principal’’ from ‘‘principal component’’ to describe independent features that are linear combinations of many other features.





**Principal feature 1 – branch points**

Each vascular network can be modeled by a mathematical tree, known as an undirected graph. There are two types of nodes on a vascular network – branch node and end node. As depicted in Fig. 4a, a branch node is where a vessel splits into multiple branches and an end point/node is a terminal point on the network.

Let  $x = NumEndPoints$  be the total number of end nodes on the vascular network,  $y = NumBranchPoints$  be the total number of branch points found in the network,

and  $z = MurrayBranchesUsed$  be the total number of branches that have child branches. Then  $z$  can be obtained by taking the difference between the total number of branches and  $x$ . If we further let  $n =$  number of nodes in the network, then the total number of branches is  $n - 1$ . Overall, we have  $n = x + y$ . Therefore,

$$x = n - y \quad \text{and} \quad z = n - 1 - x = n - 1 - (n - y) = y - 1.$$

This is to say, all three vascular attributes with significant weights in PC1 are functions of  $y$ ,  $NumBranchPoints$ .

**Principal feature 2 – diameter/thickness**

Vascular networks are intrinsically 3-dimensional tubular structure that can be modeled by circular cylinders. In this study, the diameter, which is a 3-dimensional feature of the vascular tubes, is treated the same as the width/thickness of the rectangular region obtained when tubes are pressed down, as depicted in Fig. 4b. The “pressed-down” effect is similar to that of a stereographic projection. With this in mind,  $MeanThickness$  measures the average thickness among all arteries, i.e.,

$$MeanThickness = \frac{1}{T} \sum_{i=1}^T d_i,$$

where  $d_i$  is the thickness of the  $i^{th}$  arterial vessel and  $T$  is the total number of arterial vessels exhibited in a single placental arterial network.  $StdThickness$  measures the standard deviation of thickness among all arterial vessels, i.e.,

$$StdThickness = \sqrt{\frac{1}{T} \sum_{i=1}^T (d_i - \bar{d})^2},$$

where  $\bar{d}$  is the  $MeanThickness$ .  $Volume$  gives the sum of all arterial volumes, i.e.,

$$Volume = \sum_{i=1}^T \pi \left(\frac{d_i}{2}\right)^2 \cdot c_i,$$

where  $c_i$  is the arc length of the  $i^{th}$  artery. All three features are functions of individual artery thickness and independent from the first principal feature.

**Principal feature 3 – tortuosity**

Tortuosity is a measure for the amount of twist or turns a curve has. It can be defined, in its simplest form, as the ratio of the length of the curve ( $c$ ) to the distance between the ends of it ( $d$ ), i.e.,

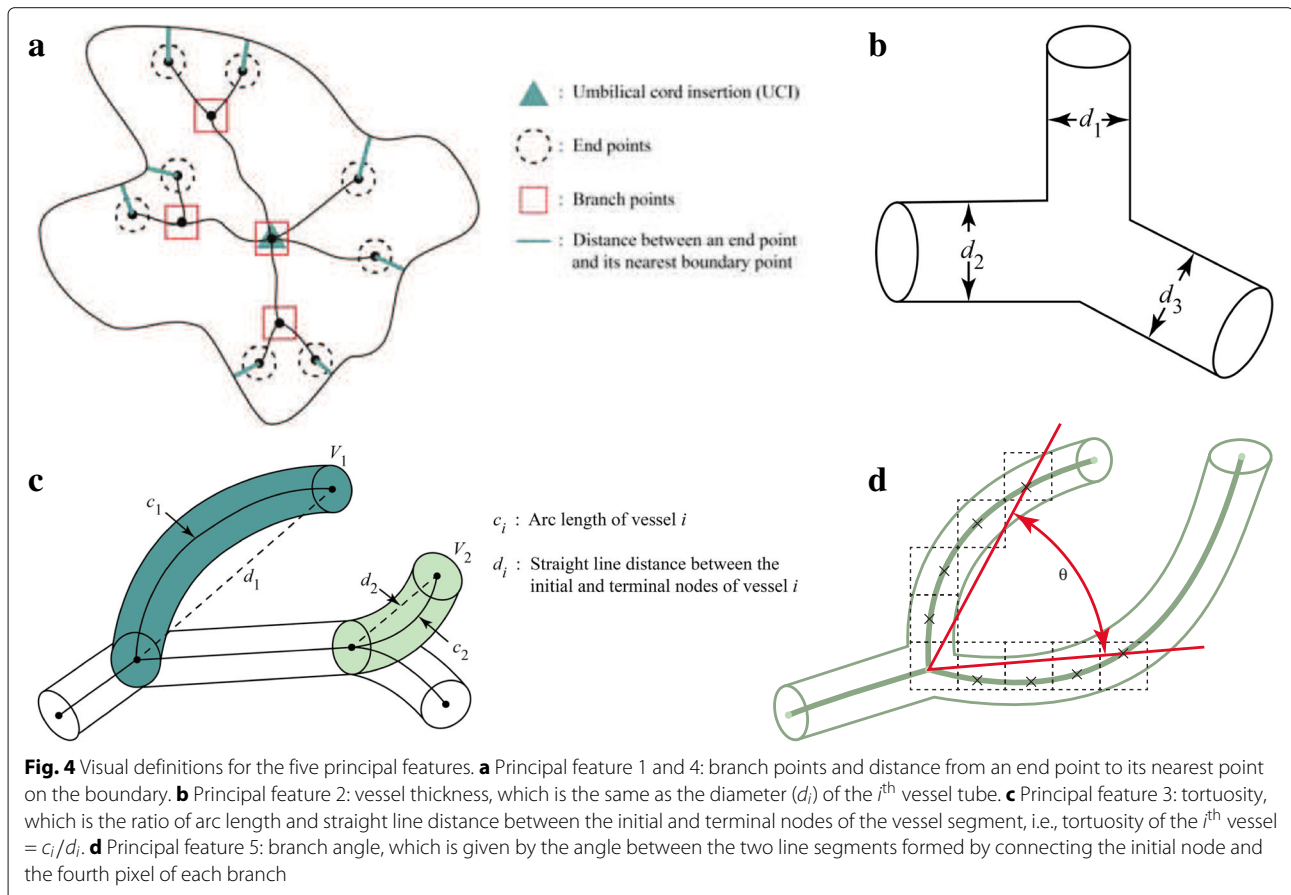
$$Tortuosity \text{ of the } i^{th} \text{ vessel} = \frac{c_i}{d_i},$$

as depicted in Fig. 4c. Severe tortuosity in vasculature can lead to various serious symptoms [23]. For example, tortuous artery and veins have been linked to aging,

**Table 1** The first five principal components (PCs) of the data retain approximately 88% of the data variability

Boruta ranking	Vascular features (variability captured)	PC1 (35.27%)	PC2 (22.57%)	PC3 (17.20%)	PC4 (7.79%)	PC5 (5.80%)
1	MeanThickness	-0.1582	-0.4747	0.1035	0.0651	-0.0089
2	MeanTortuosity	0.0002	0.0575	0.5347	-0.0979	0.0013
3	MurrayL1FitError	-0.256	-0.3903	0.0438	0.0139	0.0397
4	StdThickness	-0.1566	-0.4762	0.0701	-0.0046	0.0196
5	StdDevTortuosity	0.0029	0.0812	0.5912	-0.0641	0.1449
6	MaxTortuosity	0.0948	0.0724	0.5459	-0.0264	0.1709
7	MeanAngle	-0.0611	0.0704	0.2028	0.2135	-0.936
8	NumEndPoints	0.4251	-0.0298	-0.0132	0.0153	-0.005
9	ArcLength	0.3773	-0.1259	-0.0035	-0.0163	0.0116
10	NumBranchPoints	0.4254	-0.0301	-0.0125	0.0146	-0.0038
11	MurrayBranchesUsed	0.4254	-0.0301	-0.0125	0.0146	-0.0038
12	Volume	0.1444	-0.4823	0.065	0.0502	-0.0368
13	NumGenerations	0.3182	-0.0237	0.014	0.2178	-0.0619
14	MeanDistEndPointToPerim	0.0055	-0.0323	0.0545	0.905	0.2124
15	VesselToDiscPercent	0.255	-0.3502	0.0031	-0.2561	-0.1457

The absolute value of the attributes within each PC gives a measure of contribution. The higher the value, the bigger the contribution. Specifically, *NumEndPoints*, *NumBranchPoints*, and *MurrayBranchesUsed* contributed the most to PC1, *Thickness*, *StdThickness*, and *Volume* contributed the most to PC2, *MeanTortuosity*, *StdDevTortuosity*, *MaxTortuosity* contributed the most to PC3, *MeanDistEndPointToPerim* contributed most to PC4, and *MeanAngle* contributed most to PC5



atherosclerosis, hypertension, genetic defects and diabetes mellitus in clinical settings [24–28].

With this definition, *MeanTortuosity*, *StdDevTortuosity*, and *MaxTortuosity* give the mean, standard deviation, and maximum of the arterial tortuosities. All three variables are estimators of network’s tortuosity which is independent from the number of branch points the network has and vessel thickness.

**Principal feature 4 – growth extension**

The bolded lines in Fig. 4a illustrates the way we define the distance from an end point of the arterial network to its nearest point on the chorionic plate boundary. *MeanDistEndPtToPerim* represents the average distance between end points and their nearest point on the placental chorionic surface boundary, i.e.,

$$MeanDistEndPtToPerim = \frac{1}{m} \sum_{i=1}^m d_i,$$

where *m* is the total number of end points in the arterial network and

$$d_i = \min_{y \in \Omega} \|x_i - y\|$$

is the distance between each arterial end node,  $x_i$ , and the nearest point *y* in the boundary curve,  $\Omega$ .

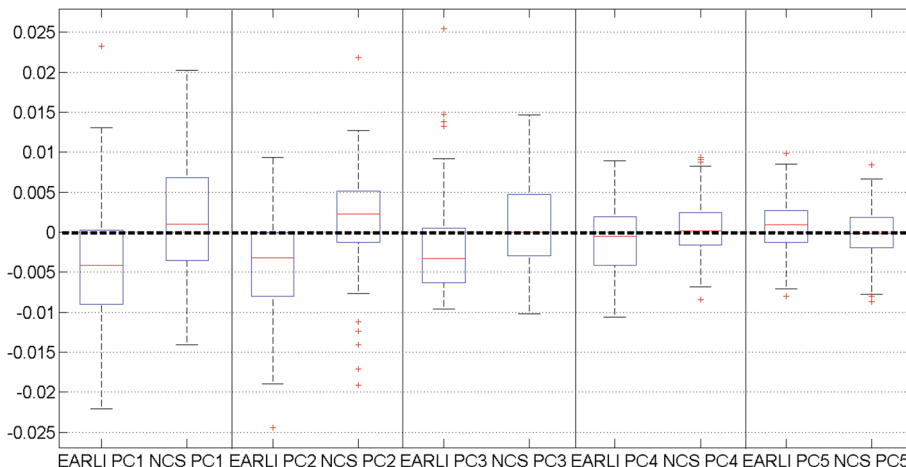
*MeanDistEndPtToPerim* gives a notion of growth extension; that is, the smaller this value is, the more extended the network is to its boundary, on average. This measure is clearly distinct from principal features 1, 2, and 3 since there is no way we can calculate this value based on existing knowledge of the previous three.

**Principal feature 5 – branch angle**

Branch angles are used to capture the *instantaneous* growth at each branch point. For simplicity, we only consider vessels that bifurcate at their respective branch point, which make up more than 90% of the data. As illustrated in Fig. 4d, branch angle is calculated as the angle between line segments that are formed between the branch point and the fourth pixel on the respective branch. The choice of four is an empirical decision to mimic the effect of instantaneous change. *MeanAngle* gives the average of all arterial vessels’ branch angles and does not depend on any of the previous four principal features. An alternative and popular approach to calculate branching angle is the one that finds the angle between two line segments joining the branch node and the end node. With this definition, branches that start off far apart but end up colliding at a single node would have an angle of 0. This alternative notion of the branching angle does not accurately capture the instantaneous growth behavior at branch points; hence, not ideal in our analysis.

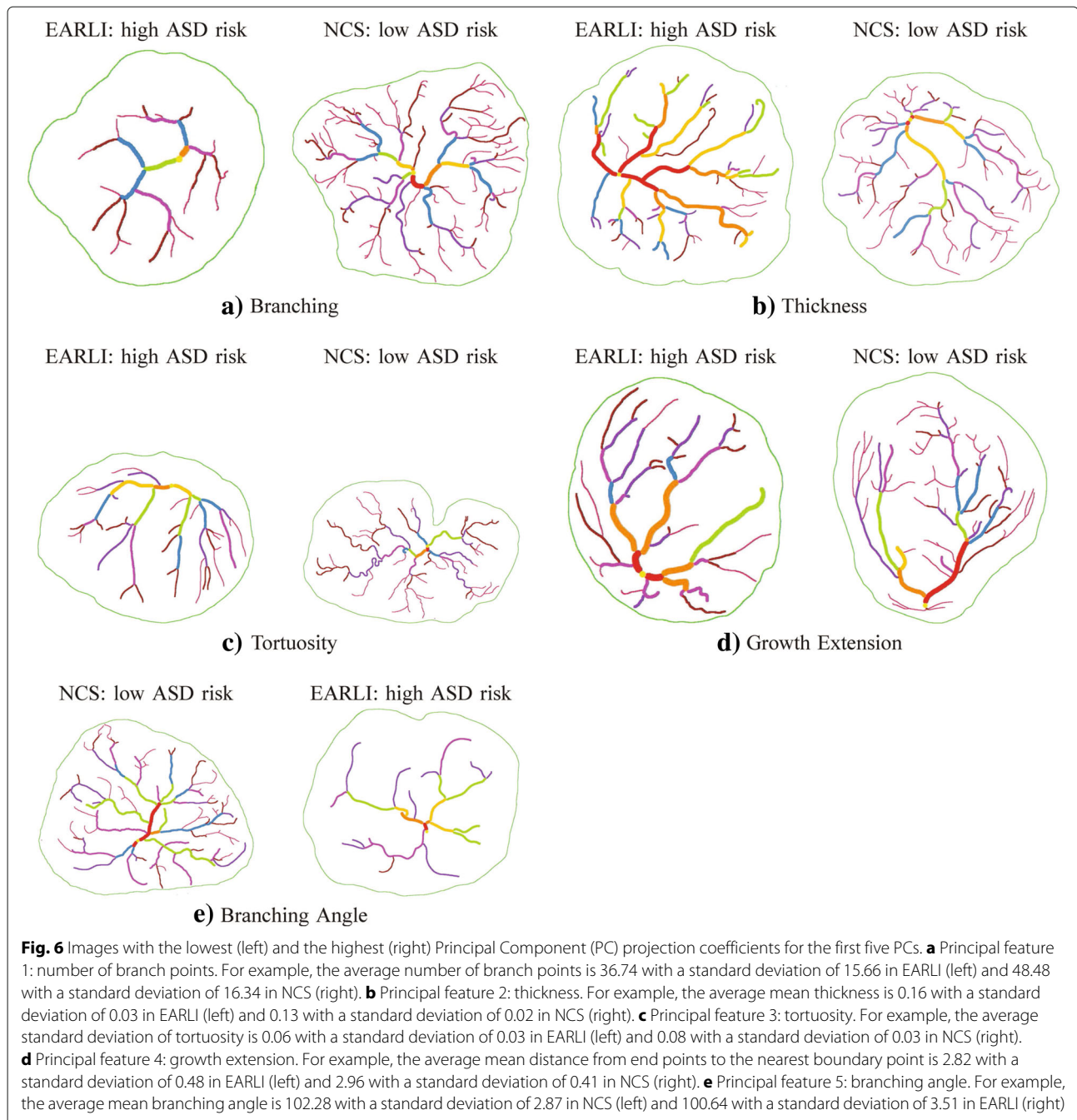
**Visualization of high- and low-risk ASD cohorts**

The numerical distribution of the high- and low-risk ASD placentas in each of the five principal features can be seen in Fig. 5. The difference between the two groups were particularly pronounced in the number of branch points, vessel thickness, and vessel tortuosity. These differences can be visualized more clearly when we compare the most extreme cases within each principal feature, as illustrated in Fig. 6. For example, the average number of branch points in the EARLI placentas was a lot lower than that in NCS, as illustrated by Fig. 6a. Similarly, a significant



**Fig. 5** Visible difference between high- and low-risk ASD groups in low dimensions. The box whisker plot of the projection coefficients for the first five principal components of EARLI (89 data points) and NCS (201 data points) cohorts. The difference between the two groups are apparent and consistent across all five PCs. For example, the mean of the first PC projection coefficients among the EARLI placentas is negative while the mean of the first PC projection coefficients among the NCS placentas is positive





difference between the two groups was found in each of the other four principal features as shown in Fig. 6.

#### Classification result of the high-risk ASD placentas

LDA with a 10-fold cross validation (CV), implemented in MATLAB, was performed to examine how well the selected principal features work together to classify placentas with increased ASD risk. The average error rates across all 10 validation trials were 6.90% and 8.97% for false positives and false negatives, respectively. The

results suggested that on average, we were able to correctly tell whether a given placenta belongs to a low-risk or high-risk ASD cohort 84 out of 100 times based on various constructs of the five extracted principal arterial features. Among the ones missed, roughly 9% were EARLI placentas misclassified as NCS placentas and 7% were NCS placentas misclassified as EARLI placentas.

To increase the reliability of our results, we additionally implemented a stratified 10-fold cross validation to take

into consideration of the class imbalance in the data set. The result was comparable to our original CV result with an overall misclassification of 15.12%.

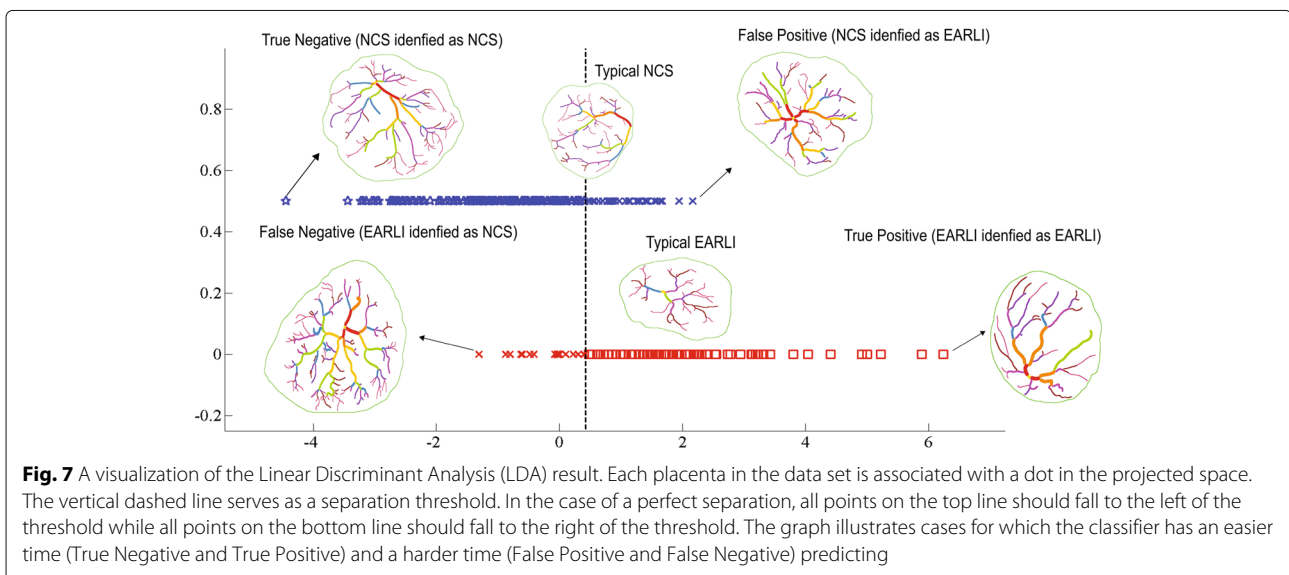
**Discussion**

A major contribution of our work is the creation and validation of a model to classify placentas associated with children in a high-risk ASD group against a population of unknown ASD risk based on automatically selected PCSVN features. The feature-selection algorithm that is based on the Boruta method returned 15 ranked attributes in an ensemble of 28 arterial features and 8 shape-related features. The fact that the Boruta method ranked arterial features higher than all of the shape-related features tells us that the difference in ASD risk can be explained by arterial features alone. We benchmarked our results with another feature selection method called Elastic Net [29], which is also known to minimize over-fitting issues. The Elastic Net method returned a set of 16 features. Among the 16 features, 14 were identical to the Boruta result with two new features, *MedianAngle* and *Kurtosis-Tortuosity*. The feature *VesselToDiskPercent* was present in the Boruta method only. We then conducted a Principal Component Analysis on the set and noticed that the principal features selected were, sorted by the amount of variance captured, (1) number of branch points, (2) tortuosity, (3) thickness, (4) branching angle, and (5) growth extension. Notice that the types of principal features selected by the Elastic Net and PCA combination are identical to those selected from the Boruta and PCA combination. The only difference is in the amount of the variability each feature captures.

The statistical significance of our results was established through the Linear Discriminant Analysis with a

10-fold cross validation. Specifically, our classifier trained on the five PCA-reduced principal features placed unlabeled placentas in the correct group nearly 84% of the time. We were able to improve the overall classification rate to slightly above 90% with a non-linear classifier called support vector machine (SVM). As mentioned earlier, since NCS was population-based, one would expect that some small number of the pregnancies resulted in a child with ASD and would thus have been “high-risk.” Therefore, a perfect classification result was unlikely. The improved classification rate afforded by SVM might therefore be a result of over-fitting. The misclassification result returned by LDA can be visualized in Fig. 7. Specifically, when a high-risk ASD placenta had too many branch points, thinner and tortuous arterial vessels, larger branching angles, and did not extend closely to the surface boundary, it was treated as if it were a low-risk ASD placenta.

Research [7] has shown that about 20% of the high-risk group will go on to have ASD, compared to roughly 1.5% in the low-risk group. Among the ASD high-risk group, 30–40% will have other developmental delays, compared to roughly 5–15% in the low-risk group. That is, the remaining 40–50% of the ASD high-risk placentas will be typically developing. For this reason, one should expect that the collections of PCSVN signatures selected by Boruta and PCA for the high-risk ASD and the diagnosed ASD placentas will not overlap completely. We made no attempt to differentiate the vascular features of the placentas associated with ASD and those associated with other developmental delays since our data does not come with such diagnoses. Our study offers no additional insight into what could have caused the placentas associated with the high ASD risk to grow this way. That is, it remains



unclear what PCSVN characteristics is unique to ASD placentas.

Some interesting questions to ask next include what environmental or genetic factors cause this group of five parameters to vary together and whether these variables stabilize in their permanent state early in gestation. Furthermore, searching for the types of geometric signatures that are measurable and capable of providing accurate readings in 3-dimensional imaging environment is also going to play a vital role in early risk assessment and intervention for ASD.

Many important ultimate placental morphologic features are likely predetermined early in pregnancy. Reliable quantization of PCSVN features will provide researchers useful tools to study the intrauterine origins of diverse disease outcomes and lead to the development of methods more broadly applicable to other branched structures including other vascular networks. Improved understanding of the details of early placental development as expressed in PCSVN branching morphogenesis may shed light on the interplay between the fetal genetic program and intrauterine environmental factors that may vary across gestation [30, 31]. Because the placenta is key to the development of many fetal/perinatal/neonatal and potentially lifelong health risks, the work presented here helps to translate placental research that can clarify timing and nature of conceptus compromise into potentially actionable clinical risk assessment.

The study presented here should motivate a pursuit of additional PCSVN features which might be correlated with various dichotomous health outcomes as long as information on outcome classification is available. We anticipate that some PCSVN features will correlate with outcomes such as diabetes, obesity, hypertension and cardiovascular disease or other “fetal origins” disorders, including autism and schizophrenia, once reliable and automated vessel extraction methods are established to allow analysis of PCSVNs in large cohorts.

Since digital images of PCSVN can be captured and analyzed within days of delivery, our classification model allows us to determine, within minutes, which risk group a new placenta belongs to. This information can be one of the multiple measures doctors use to make recommendations for early ASD interventions in clinical settings. However, a major barrier in implementing the results of our work in clinical practices is the availability of trained human tracers. Tracing PCSVN is the most time-consuming and laborious step in the entire classification pipeline. Researchers are currently developing reliable methods to automatically extract placental vascular networks from digital images of placental chorionic surface [32, 33] in order to bypass the need for manually traced images.

## Conclusions

Our study specifically demonstrated that the arterial networks that are associated with a high risk for ASD tend to have a fewer number of branch points, thicker and less tortuous vessels, better extension to the surface boundary, and smaller branch angles than their population-based counterparts. These five independent geometric features work collectively to provide a discriminating vascular signature for the high-risk ASD placentas. This result does not imply that all high-risk ASD children will have placentas satisfying each of those five conditions simultaneously; rather, these features, when taken as a whole, provide substantive discriminatory power.

The combination of the feature-selection and classification algorithms presented herein provides a mechanism in discriminating placentas from high-risk ASD pregnancies against those from a population-based cohort with unknown risks based on automatically selected PCSVN features. Although our study which built upon a single risk cohort can only offer limited implications, our work is readily transferrable to studying other adult and neonatal diseases. We will be in a great position to conduct a comprehensive study across many disease cohorts as soon as the data becomes available.

## Abbreviations

ASD: Autism spectrum disorder; EARLI: Early autism risk longitudinal investigation; LDA: Linear discriminant analysis; MZSA: Maximum z-score among all shadow attributes; NCS: National children's study; PCA: Principal component analysis; PC: Principal component; PCSVN: Placental chorionic surface vascular network; SVM: Support vector machine

## Acknowledgements

The authors would like to thank Ryan de Vera for suggesting the use of Boruta algorithm in obtaining relevant PCSVN features. The authors also wish to thank the following people who contributed to the collection of the placentas in the National Children's Study Placenta Consortium: CJ Stodgell, L Salamone, LI Ruffolo, A Penmetsa, P Weidenborner (University of Rochester), J Culhane, S Wadlinger, M Pacholski, MA Kent, L Green (University of Pennsylvania), R Wapner, C Torres, J Perou (Columbia University), P Landrigan, J Chen, L Lambertini, L Littman, P Sheffield, A Golden, J Gilbert, C Lendor, S Allen, K Mantilla, Y Ma (Ichan School of Medicine), S Leuthner, S Szabo (Medical College of Wisconsin), JL Dalton, D Misra (Placenta Analytics), N Thieck, K Gutzman, A Martin, B Specker (South Dakota University), J Swanson, C Holliday, J Butler (University of California at Irvine), A Li, RMAP S Dassanayake, J Nanes, Y Xia (University of Illinois at Chicago), JC Murray, TD Busch, J Rigdon (University of Iowa), Kjersti Aagaard, A Harris (Baylor College of Medicine), TH Darrach, E Campbell (Boston University), N Dole, J Thorp, B Eucker, C Bell (University of North Carolina at Chapel Hill), EB Clark, MW Varner, E Taggart, J Billy, S Stradling, J Leavitt, W Bell, S Waterfall (University of Utah), B O'Brien, M Layton, D Todd, K Wilson, MS Durkin, M-N Sandoval (Westat, Inc).

## Funding

None.

## Availability of data and materials

The supporting data sets analyzed during the current study are available from the corresponding author on a reasonable request.

## Authors' contributions

JMC conceived of the project aims and design. JMC, YC, ZH, and RH completed the exploration and analysis of the classification techniques applied to the data set. JMC, YC, ZH, RH, and RS implemented codes in

MATLAB, R, and SAS for analysis. JMC composed and edited drafts of the manuscript, all other authors provided feedback for revisions. CS, CN, RM, PK, JM, MF, CW, and LC were instrumental in the data collection stage of this work. All authors directly participated in the planning, execution, or analysis of the study. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Written informed consent regarding the NCS data set was obtained from all participants, and all procedures involving human subjects were approved by the Institutional Review Board of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). Written informed consent regarding the EARLI data set was obtained from all participants and/or their parent/guardian in accordance with the Drexel University Institutional Review Board approved protocol. This study concerns with secondary analysis on de-identified data; local ethics committees ruled that no formal ethics approval was required in this particular case.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Mathematics and Statistics, California State University, Long Beach, CA 90840-1001 Long Beach, USA. <sup>2</sup>Department of Statistics, Tamkang University, No.151, Yingzhuang Rd., 25137 New Taipei City, Taiwan. <sup>3</sup>Placental Analytics, LLC, New Rochelle, NY, USA. <sup>4</sup>Institute for Basic Research, Staten Island, NY, USA. <sup>5</sup>NIH National Children's Study Placenta Consortium, Bethesda, MD, USA. <sup>6</sup>Drexel University, Philadelphia, PA, USA. <sup>7</sup>University of Rochester, Rochester, NY, USA. <sup>8</sup>NICHD, Bethesda, MD, USA. <sup>9</sup>Johns Hopkins University, Baltimore, MD, USA. <sup>10</sup>University of California Davis, Davis, CA, USA. <sup>11</sup>Kaiser Permanente Division of Research, Oakland, CA, USA.

Received: 22 April 2017 Accepted: 21 November 2017

Published online: 06 December 2017

#### References

- Christensen DL, Baio J, Braun KVN, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2012. *MMWR Surveill Summ* 2016. 2016;65:1–23. doi:10.15585/mmwr.ss6503a1.
- Buescher AV, Cidav Z, Knapp M, Mandell DS. Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA Pediatr*. 2014;168(8):721–8. doi:10.1001/jamapediatrics.2014.210.
- Rosenberg RE, Law JK, Yenokyan G, McGready J, Kaufmann WE, Law PA. Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch Pediatr Adolesc Med*. 2009;163(10):907–14.
- Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J, Fedele A, Collins J, Smith K, Lotspeich L, Croen LA, Ozonoff S, Lajonchere C, Grether JK, Risch N. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatr*. 2011;68(11):1095–102.
- Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, Bryson S, Carver LJ, Constantino JN, Dobkins K, Hutman T, Iverson JM, Landa R, Rogers SJ, Sigman M, Stone WL. Recurrence risk for autism spectrum disorders: A baby siblings research consortium study. *Pediatrics*. 2011;128:488–95.
- Sumi S, Taniai H, Miyachi T, Tanemura M. Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in nagoya, Japan. *J Hum Genet*. 2006;51:518–22.
- Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, Bryson S, Carver LJ, Constantino JN, Dobkins K, Hutman T, Iverson JM, Landa R, Rogers SJ, Sigman M, Stone WL. Recurrence risk for autism spectrum disorders: a baby siblings research consortium study. *Pediatrics*. 2011;128:488–95. doi:10.1542/peds.2010-2825.
- Davies JA. Do different branching epithelia use a conserved developmental mechanism? *Bioessays*. 2002;24:937–48. doi:10.1002/bies.10161.
- Leach L, Taylor A, Sciota F. Vascular dysfunction in the diabetic placenta: causes and consequences. *J Anat*. 2009;215:69–76. doi:10.1111/j.1469-7580.2009.01098.x.
- Herr F, Baal N, Zygmunt M. Studies of placental vasculogenesis: a way to understand pregnancy pathology? *Z Geburtshilfe Neonatol*. 2009;213:96–100. doi:10.1055/s-0029-1224141.
- Barut F, Barut A, Gun BD, Kandemir NO, Harma MI, Harma M, Aktunc E, Ozdamar SO. Intrauterine growth restriction and placental angiogenesis. *Diagn Pathol*. 2010;22:5–24. doi:10.1186/1746-1596-5-24.
- Yezzi A, Kichenassamy S, Kumar A, Oliver P, Tannenbaum A. A geometric snake model for segmentation of medical imagery. *IEEE Trans Med Imaging*. 1997;16:199–209. doi:10.1109/42.563665.
- Chang JM, Mulgrew A, Salafia C. Characterizing placental surface shape with a high-dimensional shape descriptor. *Appl Math*. 2012;3:954–68. doi:10.4236/am.2012.39143.
- Shah RG, Salafia CM, Girardi T, Conrad L, Keaty K, Bartleotc A. Shape matching algorithm to validate the tracing protocol of placental chorionic surface vessel networks. *Placenta*. 2015;36:944–6. doi:10.1016/j.placenta.2015.05.004.
- Ottman R. Gene-environment interaction: definitions and study designs. *Prev Med*. 1996;25:764–70. doi:10.1038/npre.2008.2653.1.
- Yampolsky M, Salafia C, Shlakhter O, Haas D, Eucker B, Thorp J. Centrality of the umbilical cord insertion in a human placenta influences the placental efficiency. *Placenta*. 2009;30:1058–64. doi:10.1016/j.placenta.2009.10.001.
- Salafia C. Placental vascular tree as biomarker of autism/ASD risk. Annual report for U.S. Army Medical Research and Materiel Command at Fort Detrick, Maryland 21702-5012 W81XWH-10-1-0626, Research Foundation for Mental Hygiene 2014. <http://www.dtic.mil/dtic/tr/fulltext/u2/a575079.pdf>. Accessed 1 Apr 2017.
- Newschaffer CJ, Croen LA, Fallin MD, Hertz-Picciotto I, Nguyen DV, Lee NL, Bery CA, Farzadegan H, Hess HN, Landa RJ, Levy SE, Massolo ML, Meyerer SC, Mohammed SM, Oliver MC, Ozonoff S, Pandey J, Schroeder A, Shedd-Wise KM. Infant siblings and the investigation of autism risk factors. *J Neurodev Disord*. 2012;4:1–16. doi:10.1186/1866-1955-4-7.
- Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
- Kursa M, Rudnicki W. Feature selection with the boruta package. *J Stat Softw*. 2010;36:1–13. doi:10.18637/jss.v036.i11.
- Kursa M, Jankowski A, Rudnicki W. Boruta – a system for feature selection. *Fundam Informaticae*. 2010;101:271–85.
- Kursa M, Rudnicki W. The all relevant feature selection using random forest. *CoRR*. 2011; abs/1106.5112. <https://arxiv.org/abs/1106.5112>. Accessed 28 Nov 2017.
- Han HC. Twisted blood vessels: symptoms, etiology and biomechanical mechanisms. *J Vasc Res*. 2012;49:185–97. doi:10.1159/000335123.
- Corso LD, Moruzzo D, Conte B, Agelli M, Romanelli AM, Pastine F, Protti M, Pentimone F, Baggiani G. Tortuosity, kinking, and coiling of the carotid artery: expression of atherosclerosis or aging? *Angiology*. 1998;49:361–71. doi:10.1177/000331979804900505.
- Hiroki M, Miyashita K, Oda M. Tortuosity of the white matter medullary arterioles is related to the severity of hypertension. *Cerebrovasc Dis*. 2002;13:242–50. doi:10.1159/000057850.
- Pancera P, Ribul M, Presciuttini B, Lechi A. Prevalence of carotid artery kinking in 590 consecutive subjects evaluated by echo-color doppler. is there a correlation with arterial hypertension? *J Intern Med*. 2000;248:7–12. doi:10.1046/j.1365-2796.2000.00611.x.
- Callewaert BL, Willaert A, Kerstjens-Frederikse WS, Backer JD, Devriendt K, Albrecht B, Ramos-Arroyo MA, Doco-Fenzy M, Hennekam RC, Peyerit RE, Krogmann ON, Gillissen-Kaesbach G, Wakeling EL, Nik-Zainal S, Francannet C, Maurant R, Booth C, Barrow M, Dekens R, Loeys BL, Coucke PJ, Paepe AMD. Arterial tortuosity syndrome: clinical and molecular findings in 12 newly identified families. *Hum Mutat*. 2008;29:150–8. doi:10.1002/humu.20623.
- Owen CG, Newsom RS, Rudnicka AR, Barman SA, Woodward EG, Ellis TJ. Diabetes and the tortuosity of vessels of the bulbar conjunctiva. *Ophthalmology*. 2008;115:27–32. doi:10.1016/j.ophtha.2008.02.009.
- Zou H, Trevor H. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301–20.

30. Salafia C, Yampolsky M, Shlakhter A, Mandel D, Schwartz N. Variety in placental shape: when does it originate? *Placenta*. 2012;33:164–70. doi:10.1016/j.placenta.2011.12.002.
31. Schwartz N, Mandel D, Shlakhter O, Coletta J, Pessel C, Timor-Tritsch I, Salafia C. Placental morphologic features and chorionic surface vasculature at term are highly correlated with 3-dimensional sonographic measurements at 11 to 14 weeks. *J Ultrasound Med*. 2011;30:1171–8.
32. Chang JM, Huynh N, Vazquez M, Salafia C. Vessel enhancement with multiscale and curvilinear filter matching for placenta images. In: *Proceedings of the 2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2013. p. 125–128. doi:10.1109/IWSSIP.2013.6623469.
33. Yacoubou Djima Y, Salafia C, Miller RK, Wood R, Katzman PJ, Stodgell C, Chang JM. Enhancing placental chorionic surface vasculature from barium-perfused images with directional and multiscale methods. *Placenta*. 2017;57:292–3.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

