Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference
IDETC/CIE 2018
August 26-29, 2018, Quebec City, Quebec, Canada

DETC2018-85274

DESIGN SPACE EXPLORATION IN SPARSE, MIXED CONTINUOUS/DISCRETE SPACES VIA SYNTHETICALLY ENHANCED CLASSIFICATION

Tyler Wiest
tylerwiest@utexas.edu
Mechanical Engineering Department
The University of Texas at Austin
Austin, Texas, USA

Carolyn Conner Seepersad
ccseepersad@mail.utexas.edu
Mechanical Engineering Department
The University of Texas at Austin
Austin, Texas, USA

Michael Haberman
haberman@utexas.edu
Mechanical Engineering Department
The University of Texas at Austin
Austin, Texas, USA

ABSTRACT

Exploration of a design space is the first step in identifying sets of high-performing solutions to complex engineering problems. For this purpose, Bayesian network classifiers (BNCs) have been shown to be effective for mapping regions of interest in the design space, even when those regions of interest exhibit complex topologies. However, identifying sets of desirable solutions can be difficult with a BNC when attempting to map a space where high-performance designs are spread sparsely among a disproportionately large number of low-performance designs, resulting in an imbalanced classifier. In this paper, a method is presented that utilizes probabilities of class membership for known training points, combined with interpolation between those points, to generate synthetic high-performance points in a design space. By adding synthetic design points into the BNC training set, a designer can rebalance an imbalanced classifier and improve classification accuracy throughout the space. For demonstration, this approach is applied to an acoustics metamaterial design problem with a sparse design space characterized by a combination of discrete and continuous design variables.

INTRODUCTION

Design exploration is often an important part of simulation-based design. It entails acquiring new knowledge of a design space, especially the regions of the design space that are likely to lead to high-performance solutions. Exploration is often coupled with or followed by a design exploitation phase in which the emphasis is on improving or optimizing known solutions. Design exploration can be related closely to set-based design, in which the objective is to identify sets of feasible, high-performance designs rather than a single, optimal design.

Set-based design exploration entails mapping the most promising regions of the design space. A simple approach is to use intervals to capture the space, but intervals are limited in terms of accuracy and flexibility to capture complex, arbitrarily shaped regions of a design space [1-5]. Exhaustive sampling techniques have also been utilized for this purpose [6], but they can lead to prohibitive levels of computational expense. More recently, classification algorithms from machine learning have been applied for this purpose [7-11]. In previous work, Seepersad and coauthors have demonstrated how Bayesian network classifiers can be utilized and enhanced for various materials design applications [11-13] and as a basis for enhancing stochastic search for a variety of mixed discrete/continuous design problems [9], [14].

In this paper, the focus is on design exploration of a special class of mixed discrete/continuous design problems for which the promising design space is exceptionally sparse. In these cases, it is challenging to identify promising regions of the design space and even more difficult to leverage that information to further expand and improve the performance of the sparse set of candidate designs. In this paper, we address this challenge by building upon previous work in the application of Bayesian network classifiers for mapping promising regions of the design space and augment it with a synthetic oversampling technique to improve the accuracy of the classifier for these sparse design spaces.

The next section provides an overview of Bayesian network classifiers and their implementation for design spaces with continuous and discrete design variables along with an introduction to the challenges posed by sparse design spaces. Then, a synthetic oversampling method is described for addressing these challenges, followed by application to a materials design problem focused on materials with acoustic non-reciprocity.

BAYESIAN NETWORK CLASSIFIERS FOR DESIGN SPACE EXPLORATION AND MAPPING

Bayesian network classifiers have been shown to be useful in design exploration because they can be used to partition a design space according to the ability of candidate designs to meet specified performance requirements [10]. Effectively,

they enable inverse mappings of regions of interest in a design space. Furthermore, their roots in Bayesian statistics enable incorporation of prior expert knowledge and support for sequential sampling [41]. When dealing with continuous or mixed continuous/discrete design variables, kernel-based Bayesian network classifiers (KBNs) are appropriate and are the basis of the work presented here.

KBNs use Bayesian decision theory to determine the probability that a design belongs to a defined class, according to the formulation of Bayes rule in Eqn. 1.

$$P(c_l|\vec{x}) = \frac{P(\vec{x}|c_l)P(c_l)}{P(\vec{x})} = \frac{P(\vec{x}|c_l)P(c_l)}{\sum_{k=1}^{2} P(\vec{x}|c_k)P(c_k)} \quad (1)$$

In this formulation the prior probability of each class is represented by $P(c_l)$, the class conditional probability for a given set of D design variables $\vec{x} = [x_1, x_2, \dots, x_D]$ is represented by $P(\vec{x}|c_l)$, and the probability that the design belongs to a designated class is called the posterior probability of class membership and is represented by $P(c_l|\vec{x})$. Priors can be formulated in many different ways depending on the expected distributions of each class, but a simple counting prior, as shown in Eqn. 2, is often sufficient. N_l represents the number of samples in class l, while N is the total number of samples.

$$P(c_l) \cong \frac{N_l + 1}{N + 2} \quad (2)$$

To determine the class conditional probability, $P(\vec{x}|c_l)$, a kernel density estimate (KDE) is constructed. Kernel functions are centered on each candidate design point, and those functions are aggregated into the KDE. Although many kernel functions can be used for constructing KDEs, the Gaussian normal kernel is implemented in this work. Using a Gaussian KDE, the class conditional probability can be evaluated with Eqn. 3.

$$P(\vec{x}|c_l) = \frac{1}{N_l} \sum_{j=1}^{N_l} \prod_{i=1}^{D} \frac{1}{\sigma_{i,l}\sqrt{2\pi}} e^{\frac{\left(x_i - \hat{x}_i^j\right)^2}{2\sigma_{i,l}^2}} \quad (3)$$

Here, each Gaussian kernel is assigned a D-dimensional standard deviation $\vec{\sigma}$. x_i represents the design point in indicial notation and $\hat{x}_i^j$ is the data point at the center of the j^{th} kernel in the i^{th} dimension. The standard deviation sets the width of each kernel and assigning the value that yields the best performing KDE is the topic of much research [15-18], but in this work, it is treated as a heuristic parameter and tuned to maximize classification accuracy for the problem at hand. Although the variables in this discussion are described as continuous variables, discrete variables can be accommodated straightforwardly by substituting frequency-based distributions for the continuous distributions that define the class-conditional probabilities.

By applying Bayes' rule to the class conditional probabilities, the posterior probability of class membership is calculated separately for each class of interest. For example, in a binary classification scheme (e.g. high-performance, c_1, versus low-performance, c_0, with respect to specified requirements) a design is evaluated twice to determine $P(c_0|\vec{x})$ and $P(c_1|\vec{x})$, and the candidate design is assigned to be a member of the class with the larger posterior probability, according to Eqn. 4. In some cases a designer may wish to bias the assignment toward a certain class based on risk or some *a priori* knowledge. These heuristic risk factors are defined as $\lambda_l \in [0,1]$ and applied as weights on the posterior (default $\lambda_l = 1 \; \forall \; l$). The difference between the posterior probabilities is called the posterior class discriminant (PCD).

$$PCD \in [-1,1] = \lambda_1 P(c_1|\vec{x}) - \lambda_0 P(c_0|\vec{x}) = \frac{\lambda_1 P(\vec{x}|c_1)P(c_1) - \lambda_0 P(\vec{x}|c_0)P(c_0)}{P(\vec{x}|c_1)P(c_1) + P(\vec{x}|c_0)P(c_0)} \quad (4)$$

The D-dimensional hypersurfaces along which $PCD = 0$ represent decision boundaries in the space. An example of posterior probability surfaces for a binary classification in 2D is shown in Figure 1. In the figure, green and red points represent instances belonging to different classes. The blue and red surfaces represent the posterior probability of their respective classes throughout the space. The probability of each class is equal ($PCD = 0$) where these surfaces intersect as represented by the black curves in Figure 1.

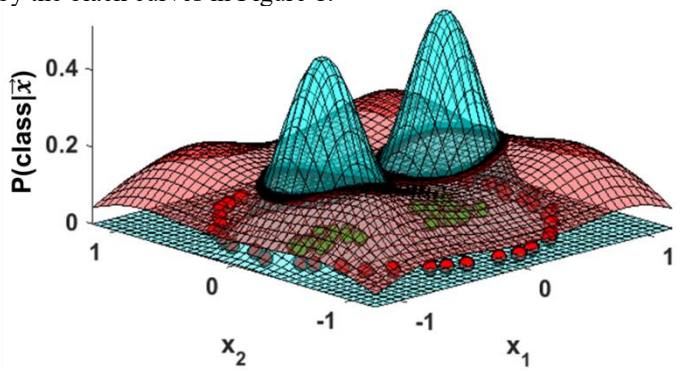

Figure 1: Posterior probability surfaces over an example 2D design space. Each surface is generated from the sample points using a kernel-based Bayesian network classifier (KBN) [12].

The hypersurfaces along which $PCD = 0$ are particularly important for accurate classification because they are the decision boundaries between design space regions of different class membership. When the number of training points available for one class is much greater than for another class, the PCD can be skewed in favor of the dominant class, a condition called imbalanced classification, which is the focus of this paper.

IMBALANCED CLASSIFICATION PROBLEMS

In many cases, analysts and designers seek to predict rare events based on existing data sets. Classic examples include identifying patients with early-stage cancer indicators from imaging data [19] in the medical field, identifying oil slicks from satellite imagery [20] or detecting instances of credit card fraud from a large number of legitimate transactions [21]. This problem is called imbalanced learning or anomaly detection in the machine learning community and is characterized by a significant class imbalance. Conventionally, the imbalanced learning problem is formulated as a binary classification in which a minority class is of particular interest and all other outcomes are grouped into a majority class. In the binary framework, a minority class instance is considered a "positive" *P* result and the majority class instance is considered a "negative" *N* outcome. With the classes defined in this way, a 2x2 confusion matrix is a convenient way to view classification performance, where TP is a true positive result, FP is a false positive, TN is a true negative, and FN is a false negative. The binary confusion matrix is shown in Figure 2 below.

		Predicted Class	
		P	N
True Class	P	TP	FN
	N	FP	TN

Figure 2: Confusion matrix for binary classification

These labels give insight into the classification task but more importantly serve as the basis for more informative evaluation metrics to make comparisons between classifiers. Quite a large number of evaluation metrics have been created for the purpose of comparing classifiers [22]. Some simple yet descriptive metrics include: true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and accuracy (ACC), as defined in Eqns. 5-8.

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)} \tag{5}$$

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)} \tag{6}$$

$$FNR = \frac{FN}{P} = \frac{FN}{(TP + FN)} = 1 - TPR \tag{7}$$

$$ACC = \frac{TP + TN}{(TP + FP + FN + TN)} \tag{8}$$

With a small number of minority class members used for training, classifiers tend to predict that all candidates belong to the majority class. When using a KBN, this occurs because the KDE for evaluating the majority class posterior probability overwhelms that of the minority class, resulting in diminished minority class regions of the design space. Since almost all instances are members of the majority class, the ACC will be very high, even if every single minority class instance is misclassified [23]. For this reason, the ACC is insufficient for evaluation of imbalanced classification tasks because identification of minority class instances is very important and misclassification can be very costly. In the cancer detection example introduced earlier, for example, misclassification of a minority class instance means that a patient with cancer is diagnosed as healthy. In a problem of this nature, classifier performance is better described by its ability to identify minority class instances, so TPR and FPR are more meaningful performance indicators for the minority class than ACC.

Due to the challenge of imbalanced classification and its prevalence in machine learning tasks [20], [24-27], significant research has focused on improving classifier performance under these conditions [23], [28-33]. The two most general approaches are to train the classifier in a cost-sensitive manner and to restore balance by resampling to either decrease the number of majority instances or increase the number of minority instances in training data [23]. Resampling the training data set can be done in many ways including (1) gathering more real samples of the minority class, (2) randomly removing majority class samples (random undersampling), (3) informed undersampling of the majority class using an algorithm to remove samples so minimal definition of the space is lost, (4) removing or altering overlapping instances from the two classes, and (5) generating synthetic samples to bolster the minority class. Each method has benefits, but the generation of synthetic minority samples has been shown to be particularly powerful both in a static data set [28] and as an adaptive sampling tool [34]. This technique—called SMOTE for Synthethic Minority Oversampling TEchnique—was introduced in a seminal paper by Chawla et al. [28]. SMOTE works by performing linear interpolation along *D*-dimensional lines between each minority class instance in the design (feature) space and its k nearest neighbors where k can be adjusted based on the desired breadth of minority oversampling. Adaptations of the original SMOTE are numerous [34-37]. The strategy described in this paper builds on SMOTE by combining it with the PCD to enhance the classifier's true positive rate (TPR).

Imbalanced classification appears in engineering design tasks in which only a small number of high-performance designs exist within a design space that contains a disproportionately large number of low-performance designs. Imbalance is exacerbated in sparse design spaces with a mix of continuous and discrete variables that cannot benefit from gradient-based optimization techniques. In these cases, adapting a SMOTE approach adds capability to improve classifier performance and inform the exploration of sparse high-performance regions within the design space.

The following section describes a novel way to generate synthetic samples of the minority class and, by using them to train a KBN, improve the accuracy of the design space mapping. As a result, the model is improved and class prediction of candidate designs becomes more accurate. This method is particularly advantageous in sparse and/or

imbalanced design spaces where the KBN model underestimates the size of high-performance regions in the design space. It uses information gained from building a KBN to intelligently select where to add synthetic training points in a design space to improve definition around decision boundaries.

POSTERIOR CLASS DETERMINANT (PCD) INFORMED SMOTE

In sparse design spaces, sampling is likely to lead to very few high-performance designs isolated locally among an overwhelming number of low-performance designs. In this case, training a kernel-based classifier yields a KDE that indicates a misleadingly small region of the space holds high-performance (minority class) designs. Figure 3 shows the effect of a class imbalance of 50:1 in a 2D design space. In this example space, a region of high performance exists near the middle of the space, but due to sampling at low density only a single high-performance instance exists in the data set. The resulting design space mapping fails to accurately represent the performance regions because of the imbalance. Predictions based on this mapping would likely misclassify any high-performing designs near the single high-performance training point because the posterior probability of class membership for the low-performance points dominates that of the high-performance point(s).

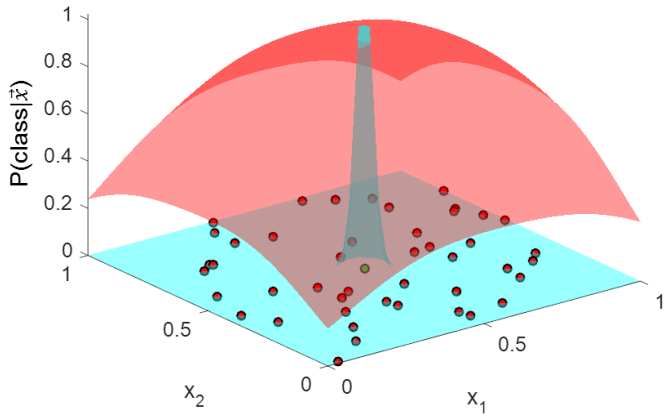

Figure 3: High-performance (blue) and low-performance (red) posterior probability surfaces over a sparse 2D design space with a class imbalance of 50:1. The 49 low-performance points are red; the single high-performance point is green. Due to imbalance, this mapping suggests that a misleadingly small region of the space holds high-performance designs.

To improve the overall classifier performance in a sparse, imbalanced design space we seek to improve the design space mapping near the decision boundaries ($PCD = 0$). Accordingly, additional high-performance training points are needed near the decision boundaries. As shown in Figure 4, the procedure starts with a sampling strategy and an initial design space mapping. According to a standard k-fold cross validation (CV) strategy, the KBN classifier is trained using a training set of candidate designs with known performance evaluated with a predictive simulation model. Then, the accuracy of the KBN is evaluated with a separate set of test data (also with known performance evaluated with the same predictive simulation model). The KBN classifies the design space into high- and low-performance regions according to performance thresholds specified by the designer. If the high-performance designs are represented sparsely in the training data, and the TPR is unacceptably low, the KBN is a candidate for synthetic sampling. Figure 4a illustrates a simplified 2D design space with one discrete variable (x_2) and one continuous variable (x_1) and sparse representation of high-performance designs, shown as green points in the figure. This KBN is a candidate for a synthetic sampling procedure because its TPR is unacceptably low, as indicated by the substantial proportion of high-performance points outside of the decision boundary, which is represented by the solid black line in Figure 4a.

The synthetic sampling procedure operates by adding synthetic training points near the decision boundaries to improve the accuracy of the classifier. The procedure begins by identifying candidate designs near a decision boundary, where $PCD = 0$ according to Eqn. 5. By specifying a PCD interval, the designer selects the design points that are suitable basis points for synthetic sampling. In Figure 4a, the decision boundary ($PCD = 0$) is represented by the solid black line, and a small interval around the PCD (e.g., $PCD \in [-0.1, 0.1]$) is represented by the dotted black line. Any points within the PCD interval are suitable basis points for synthetic sampling. The size of the PCD interval determines the extent of the design space that is utilized for synthetic sampling. An interval of $PCD \in [-1, 1]$ would encompass the entire space and an interval of $PCD \in [-0.1, 0.1]$ would cover only a small fraction of the total design space, for example.

After assigning the PCD interval, which encompasses the basis points for synthetic sampling, the next step is to interpolate between the basis points. In this case, one of the variables, x_2, is a discrete variable, which is not amenable to interpolation. Accordingly, an initial set of basis points is selected based on a common value for the discrete variable, x_2, as indicated by the red box in Figure 4a. Then, synthetic points are generated by interpolating between the basis points as shown in Figure 4b. In Figure 4b, the performance response, $f(x_1)$, is plotted as a function of the continuous variable, x_1, for the basis points. The value of the continuous variable is adjusted to generate candidate synthetic points. The performance, $f(x_1)$, of the candidate synthetic points is evaluated by interpolation between the basis points of known performance. In this case, interpolation is performed via simple linear interpolation between neighboring points, but any surrogate modeling procedure (regression, kriging, etc.) could be utilized to perform the interpolation. If the interpolated performance of the candidate synthetic point exceeds the performance threshold specified by the designer for the purposes of classification, it is accepted as a synthetic point to be added to the training set, as represented by one of the green stars in Figure 4b. If not, it is rejected, as represented by the

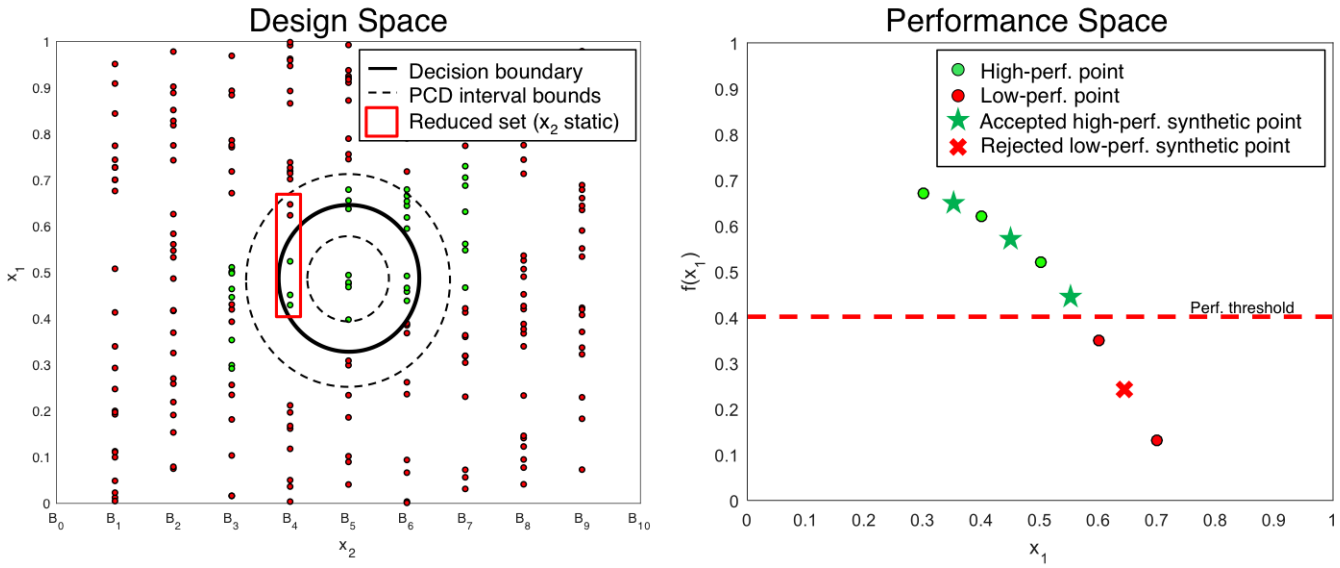

Figure 4: (a) A 2D design space with one discrete (x_2) and one continuous variable (x_1). The high-performance region is bounded by the black decision boundary derived from the KBN. As shown, several high-performance (green) points are incorrectly classified as low-performance (outside of the decision boundary), contributing to an undesirably low TPR, so a PCD interval, indicated by the dashed lines, is generated to identify basis points for synthetic sampling. (b) A linear interpolation scheme is used to generate synthetic points for the reduced set outlined by the red box in (a). If the interpolation indicates that the point is high performance, it is added to the training set as a synthetic point, as indicated by the green stars. Otherwise, it is rejected as a candidate synthetic point, as indicated by the red X.

red X in Figure 4b. Then, the process is repeated for all unique values of the discrete variable(s) until all of the basis points within the PCD interval have been considered. This simple example includes only one continuous variable, but multiple continuous variables can be accommodated via multivariable interpolation.

As shown in Figure 5, the synthetic training data is merged with the original training data to form a new training data set, and a synthetically enhanced KBN is trained. The accuracy of the synthetically enhanced KBN is evaluated with the same test data utilized to evaluate the accuracy of the original KBN. If the accuracy is still unacceptable, more synthetic training data can be generated, and the process can be repeated.

When applying this method to a design problem, it is good practice to select a subset of synthetic points to validate with the underlying simulation model. Although using surrogate models to interpolate between basis points is intended to reduce computational expense, validating a subset of synthetic points helps ensure the accuracy of the synthetically enhanced design space. The appropriate size of the validation subset primarily depends on the accuracy of the surrogate model. For example, if a linear interpolation model is used to generate and evaluate synthetic points that exhibit a highly multimodal response with few basis points, the surrogate prediction of the performance of the synthetic points could deviate significantly from the performance predicted by the underlying simulation model. In this case, it may be necessary to evaluate several of the synthetic points with the underlying simulation model. A validation step is performed in the demonstration problem in the next section.

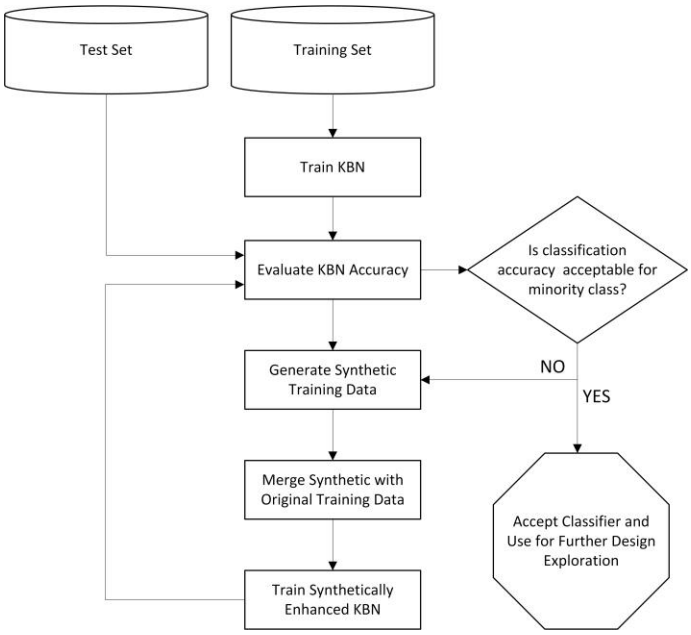

Figure 5: Flowchart outlining the strategy of PCD informed SMOTE. The steps are usually executed within a CV scheme.

DEMONSTRATION PROBLEM

To investigate the effectiveness of the PCD informed SMOTE procedure, we consider the problem of identifying acoustic non-reciprocity in simple metamaterials. A medium exhibiting acoustic non-reciprocity responds differently to identical sound waves when they are radiated into the medium from different directions. Differences in response caused by

geometric and boundary effects are generally not considered to be examples of non-reciprocity; instead, the focus is on non-reciprocity caused by the composition of the medium itself.

One kind of acoustic non-reciprocity is called Willis coupling (also referred to as bianisotropy) [38]. Analogous to magnetoelectric coupling in electromagnetism [39], Willis coupling is caused by microscale effects in a medium resulting in macroscopic field coupling between the momentum and strain constitutive relations. The coupled constitutive relations are:

$$\begin{aligned} \vec{\mu} &= \underline{\vec{\rho}} \cdot \vec{u} - \vec{\eta} p \\ \epsilon &= \vec{\gamma} \cdot \vec{u} - \beta p \end{aligned} \tag{9}$$

with coupling vectors

$$\begin{aligned} \vec{\eta} &= \vec{\chi}^o + i\vec{\chi}^e \\ \vec{\gamma} &= \vec{\chi}^o - i\vec{\chi}^e \end{aligned} \tag{10}$$

where

$\vec{\mu}$: momentum density
$\vec{u}$: particle velocity
ϵ : volume strain
p : acoustic pressure
$\underline{\vec{\rho}}$: anisotropic mass density
β : adiabatic compressibility
$\vec{\chi}^o$: odd coupling
$\vec{\chi}^e$: even coupling

A way to demonstrate this coupling is with acoustic metamaterials called Willis materials. Work by Sieck, Alu, and Haberman showed that composites can be homogenized to quantify Willis coupling on the macroscopic level [40]. The approach they presented motivates the designer to identify a set of composites that exhibit non-negligible coupling and to seek instances of strong coupling under various conditions. This problem is too broad to approach comprehensively, but by limiting the scope of the investigation, it becomes tractable. The problem is well suited to set-based design and classification of high- and low- performance designs. Additionally, it provides a challenging case of class imbalance due to the nearly infinite number of candidate designs and the relatively small fraction of those designs that meet reasonable high-performance thresholds.

For simplicity, we consider only a 1D model of a composite made from periodic unit cells. Since the composite is 1D, the waves are assumed to be plane waves. Figure 6 shows the configuration of the unit cells under plane wave radiation in the $\hat{x}$ direction. The configuration of potential unit cells is limited to one general design consisting of a two-layer inhomogeneity in a background of liquid water. The composition of the layers constituting the inhomogeneity is limited to common materials with well understood bulk material properties. In particular, density, ρ, sound speed, c, and compressibility, β, in the layers and the background water are the important factors affecting wave propagation through the unit cell. By convention the properties of the background fluid are labeled ρ_0, c_0, and β_0. Each inhomogeneity includes two layers, and each layer is assumed to have identical thickness because thickness has less effect on the wave propagation than the existence of boundaries between materials. For a unit cell of length L and an inhomogeneity of length l, the volume fraction is $VF = \frac{l}{L}$.

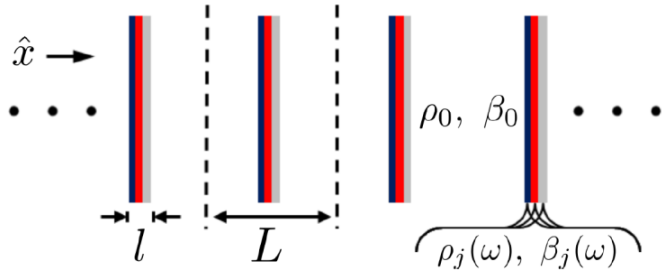

Figure 6: Plane wave propagation along the x-axis in a 1D periodic medium composed of repeating multi-layer inhomogeneities of length l in unit cells of length L with a background of liquid water with properties ρ_0, c_0, and β_0 [40].

Measurement of the Willis coupling terms $\vec{\chi}^e$and $\vec{\chi}^o$, which enter into the constitutive relationships in Eqn. 10, requires a model that captures the non-reciprocal response of the unit cell. Plane waves are radiated into the unit cell from each direction separately, and the pressure response is measured at the boundaries. Since the homogenization procedure is dynamic, the frequency of the plane waves affects the material behavior and therefore must be taken into account. For the homogenization, frequency is incorporated as an element of the wavenumber in the background material and normalized to the unit cell length. The normalized wavenumber takes the form $k_0 L = \frac{2\pi f L}{c_0}$. For quantifying the Willis coupling, we use the same normalization of the coupling terms as Sieck, Alu, and Haberman [40], namely $c_o\chi^e/k_0 L$ and $c_o\chi^o/k_0 LkL$. The normalized terms $c_o\chi^e/k_0 L$ and $c_o\chi^o/k_0 LkL$ are zero-centered and comparable for different unit cell designs. In addition to the coupling, the impedance ratio between the composite and water, $\frac{Re(Z_{eff})}{Z_0}$, is measured, along with the effective normalized wavenumber of the composite, kL. All important parameters are collected in Table 1. Note that frequency is typically an exogenous factor, but since we are interested in designing composites to perform in specific frequency ranges, it is considered a design variable by way of the normalized wavenumber.

A FEA was established to simulate the non-reciprocal response of individual unit cells. Figure 7 shows the unit cell geometry used for simulation. Simulation results are post-processed to perform dynamic homogenization and yield effective macroscopic parameters of the composite including density, compressibility, and even and odd coupling terms. The full homogenization procedure is too extensive to describe here. Interested readers may refer to Sieck, Alu, and Haberman (Sieck, Alù, & Haberman, 2017) for a detailed description.

Table 1: Key terms for use in design of 1D Willis material. A 6D design space and 4 important performance indicators are shown.

Design Variables		Performance Indicators	
Density of Layer 1	ρ_1	Normalized Even Coupling	$c_o\chi^e/k_0L$
Density of Layer 2	ρ_2	Normalized Odd Coupling	$c_o\chi^o/k_0LkL$
Sound Speed in Layer 1	c_1	Ratio of Effective Impedance to Background Impedance	$\frac{Re(Z_{eff})}{Z_0}$
Sound Speed in Layer 2	c_2		
Normalized Wavenumber in the Fluid	$k_0L = \frac{2\pi fL}{c_0}$	Effective Normalized Wavenumber	kL
Volume Fraction of Inhom.	$VF = \frac{l}{L}$		

Sampling was restricted to consider 324 possible combinations of 18 common materials (i.e. steel, rubber, glass, lead, etc.) of varying properties but equal layer thickness in a background of liquid water. A Hammersley sequence was used to uniformly sample $VF \in (0.1,\ 0.35)$ at 5 points and $f \in [500Hz,\ 50{,}000Hz]$ at 100 points in increments of 500 Hz. The authors of the Willis coupling paper suggested starting with these intervals for VF and f to find useful results [40]. With these sampling increments, the data set includes 162,000 samples.

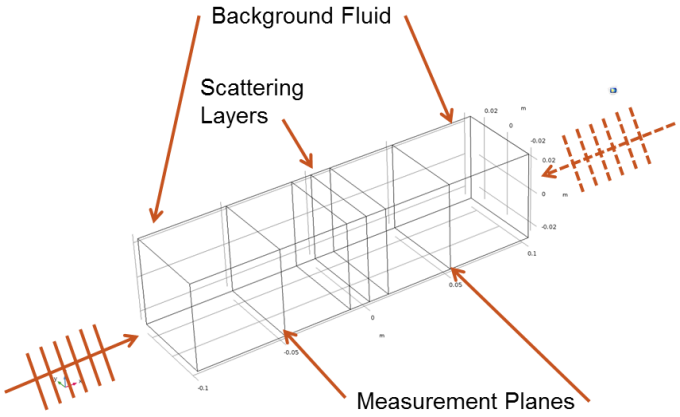

Figure 7: Simulation geometry of 2 layer inhomogeneity in background fluid

To meet performance constraints, the effective impedance of the composite was constrained to 80%-120% that of water, and the effective wavenumber was constrained to be less than π, which is considered the upper limit for the dynamic homogenization scheme to produce valid results. After removing all samples violating these constraints, the data set contained 6,750 samples in a 6D design space and was prepared for classification using KBNs. A reasonable coupling performance threshold was selected to distinguish between high- and low-performance classes. By classifying any sample

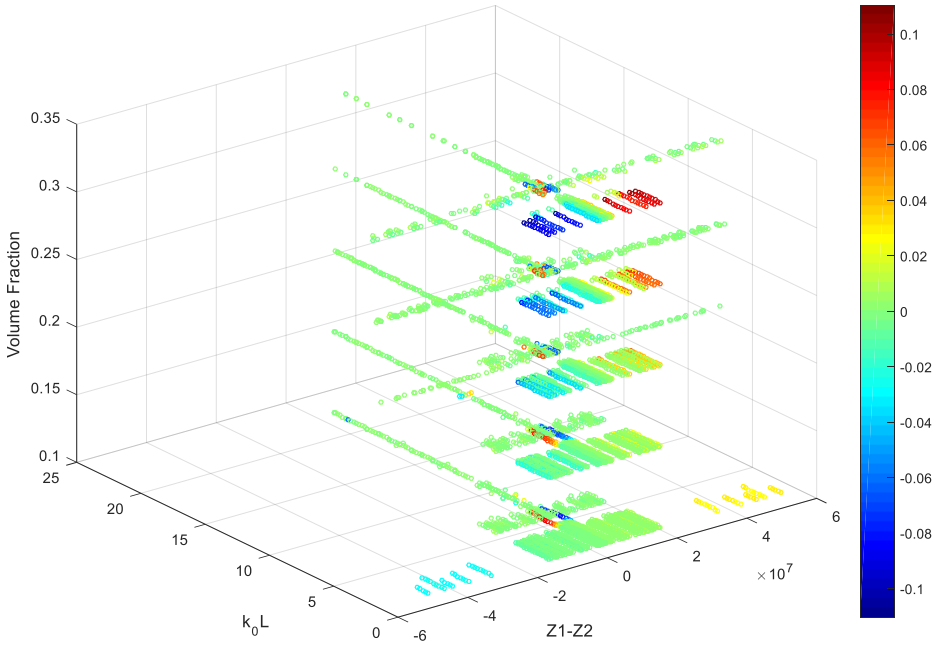

Figure 8: All sample points used for the demonstration problem scattered in a design space reduced to 3D by consolidating the material properties into a difference of characteristic acoustic impedances ($Z_1 - Z_2$). All points in this set have effective impedances within 80%-120% of water and effective wavenumbers less than π. The color map represents normalized even coupling values, with absolute values greater than 0.02 indicating high-performance.

with $|c_o\chi^e/k_0L| > 0.02$ as high-performance, 737 of the 6750 samples were classified as high-performance and 6013 samples as low-performance. For a visualization of the 6D design space, the material properties were combined by calculating the difference in characteristic acoustic impedance between the 2 layers of the inhomogeneity $Z_1 - Z_2 = \rho_1 c_1 - \rho_2 c_2$. Figure 8 shows this reduced design space in a 3D scatter plot with a color map indicating normalized even coupling, $c_o\chi^e/k_0L$.

Figure 9 shows some examples of the normalized even coupling response to changes in the normalized wave number, k_0L, for unique material combinations. The relationships are a subset of those eventually used for synthetic sample generation in the synthetically enhanced KBN model.

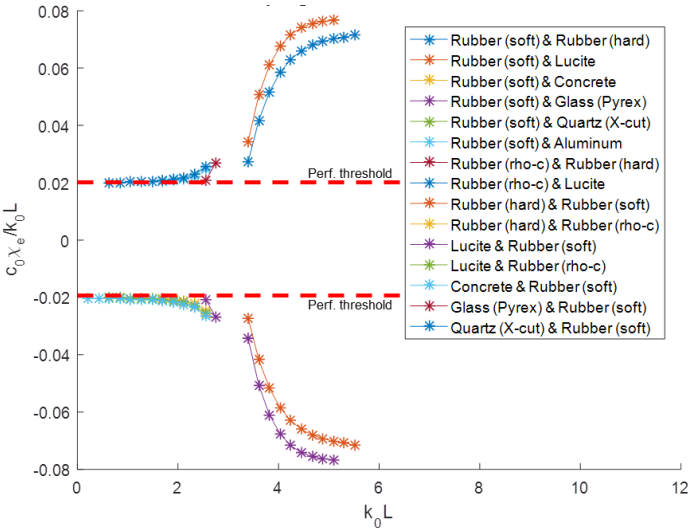

Figure 9: Normalized even coupling values as a function of normalized wave number, for 15 unique combinations of materials at a volume fraction of 17%. All responses have effective wavenumber less than π and effective impedance within 80%-120% of water.

The material properties ρ and c were considered to be discrete when sampling since they belong to distinct material choices, but they were treated as continuous design variables to generate KDEs since the problem was motivated by metamaterials that enable design for material properties on a continuous scale. Finally, this data set was used to train and cross-validate a naïve Bayes classifier both with and without using the PCD informed SMOTE method presented in this work.

The base KBN with a Gaussian prior was tuned by many iterations of cross-validation with a standard deviation formulated as

$$\sigma_{i,l} = \frac{\alpha\hat{\sigma}_{i,l}}{N_l^{1/D}} \tag{11}$$

where

$\sigma_{i,l}$: kernel width parameter (Eqn. 3)
α : heuristic scalar
$\hat{\sigma}_{i,l}$: st. dev. of design var. i for designs belonging to class l
N_l : number of samples in class l
D : number of design vars. (dimensions)

The base KBN model showed the best performance with very thin kernels (e.g. at $\alpha = 0.01$). Table 2 below summarizes the performance of the base model. Regardless of the value of α, the TPR never rose above about 38%, indicating poor KBN performance in the high-performance design space.

Table 2: Results of tuning the base KBN model by cross-validation with varying heuristic α. The highlighted entries in Table 2 match the α settings that dominate for the synthetically enhanced model.

α	TPR	FPR	ACC
0.001	0.379	0.011	0.922
0.010	0.381	0.011	0.923
0.064	0.349	0.014	0.916
0.119	0.206	0.013	0.902
0.173	0.119	0.011	0.894
0.228	0.071	0.011	0.889
0.282	0.035	0.010	0.885
0.337	0.020	0.010	0.884
0.391	0.008	0.008	0.884
0.446	0.001	0.008	0.884
0.500	0.003	0.007	0.885

The synthetically enhanced model was tuned and compared against the base KBN model by using an identical set of test data and with synthetic points added only to the training set. Tuned hyperparameters were: α, the PCD interval, and the quantity of interpolated points added between real samples. Synthetic points were generated by linearly interpolating along the curves shown in Figure 9 exactly as shown in Figure 4b.

Together, Figure 10 and Table 3 below show the results of tuning with 240 combinations of hyperparameters: $Interpolation\ Layers$: $[3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$, PCD_{upper}: $[0.1, 0.2, 0.5, 0.6, 0.8, 0.9, 0.95, 0.99]$(zero-centered), and $\alpha = [0.391, 0.446, 0.500]$. The number of interpolation layers indicates the number of synthetic points between each pair of basis points. Early tuning showed the three α values included in the 240 combinations yielded the best classifier performance and were focused upon thereafter.

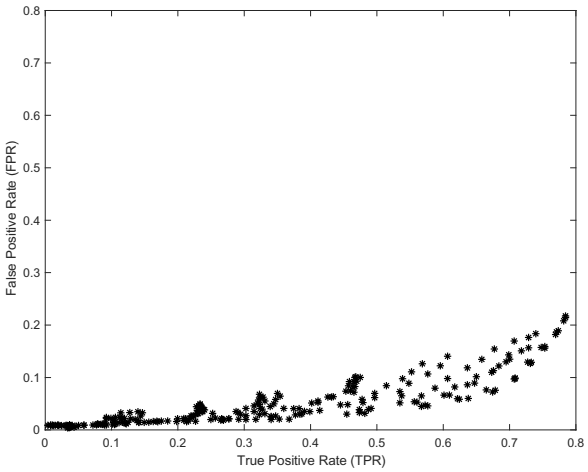

Figure 10: Performance results of tuning the synthetically enhanced model for the acoustic bianisotropy problem.

Figure 10 shows the achievable trade-off between FPR and TPR graphically for all 240 combinations while Table 3 shows the performance of 10 synthetically enhanced models with high TPR and the lowest associated FPR. It is clear that FPR rises sharply for models achieving a TPR greater than 70%. In this case, the best models are considered to be those nearest the bottom-right corner of Figure 10.

Table 3: Ten best results from tuning synthetic model sorted by TPR.

Interpolation Layers	PCD Interval	α	TPR	FPR	ACC
6	0.99	0.500	0.624	0.059	0.907
6	0.99	0.446	0.638	0.060	0.907
7	0.99	0.391	0.665	0.076	0.895
7	0.99	0.500	0.674	0.072	0.900
7	0.99	0.446	0.678	0.076	0.897
8	0.99	0.500	0.708	0.097	0.881
8	0.99	0.446	0.708	0.099	0.880
9	0.99	0.391	0.731	0.127	0.858
9	0.99	0.446	0.733	0.129	0.856
10	0.99	0.391	0.753	0.156	0.834

Increasing FPR is an undesirable result that comes along with expanding the high-performance regions of the design space mapping but is somewhat unavoidable. The ACC is slightly lower for the synthetically enhanced models than the best base model, which means that more false positives have been introduced than false negatives have been removed. However, in the interest of improving an imbalanced classifier the vast improvement in TPR outweighs the ACC reduction. With a higher TPR rate, this classifier is now much more useful for exploring a design space. Note that the large PCD intervals work well for this problem. This is likely a result of the simple relationship between the continuous variable and the performance metric and the resulting quality of the interpolation. In a design space in which the relationship is highly non-linear and multimodal, a thinner PCD interval may be necessary. Overall the performance of this classifier has been improved noticeably by adding synthetic points around the decision boundaries.

For this problem, a validation step compared the interpolated performance of the synthetic points to the results of the FEA. Plots of $c_o\chi^e/k_0L$ vs. k_0L showed mostly smooth functional relationships that were suitable for linear interpolation, but to be thorough, we evaluated 2811 synthetic design points. Simple but representative hyperparameters were chosen: $Interpolation\ Layers = 1, PCD\ Interval = [-1, 1]$, and $\alpha = 0.5$, along with a performance threshold of $|c_o\chi^e/k_0L| > 0.02$. The mean squared error between the interpolated and simulated $c_o\chi^e/k_0L$ values of the synthetic points was $7.4x10^{-6}$. This low error rate demonstrated that the interpolated performance values were very close to the result we would have gotten by incurring the computational expense to simulate them all. In this case, it required about 3.5 minutes to simulate each design point, so simulating all 2811 points required almost 170 minutes on a PC with an Intel i5 processor and 16 GB of RAM. For comparison, generating the synthetic points on the same machine takes just a few seconds.

Despite the very challenging nature of the design space, adding synthetic minority class samples to the training sets improved the classifier's performance by increasing TPR at a significantly greater rate than the FPR when heuristics were well tuned. An ideal synthetic oversampling would not increase FPR at all, but since the synthetic points are enlarging the high-performance regions of the design map, it is expected that some low-performance designs will lie in those regions and be misclassified during the cross-validation.

CONCLUSIONS AND FUTURE WORK

In this paper, a method is introduced to utilize posterior probabilities of class membership to generate synthetic points of the minority class and rebalance an imbalanced classification problem with mixed discrete/continuous variables. The method works by identifying reduced sets of basis points near the decision boundaries that partition high-performance regions of the design space and interpolating continuous design variables with discrete design variables held constant. Interpolated points that belong to the minority (high-performance) class are added to the overall training set in an attempt to artificially balance the classifier's KDEs. This method provides an advantage in design exploration by saving the computational expense of evaluating additional candidate designs when working in a sparse design space with mixed variable types.

One obvious opportunity to expand this work is to apply the method to additional problems and fully benchmark it against other approaches to the same type of problems. Another opportunity is to study the relationship between the PCD bounds and design space characteristics like sparsity on the synthetic generation process. It would also be worth considering more sophisticated interpolation techniques for

generating and evaluating synthetic training points. For example, prior knowledge of the response or transfer learning from similar problems could be incorporated to add realistic synthetic points in areas of the space with little definition.

In general, validating the classified performance of synthetic points is of great interest for improving the accuracy of the enhanced classifier, but it must be done in a cost-effective way. There is potential for developing an algorithmic scheme to validate a subset of the synthetic points to ensure that they are not misclassified. For example, a sampling scheme could be used to start evaluating synthetic points with expensive simulation models, followed by an expected improvement framework to determine which synthetic points are most valuable to evaluate with an underlying simulation model. In this way, PCD informed SMOTE would take the form of an adaptive sampling technique.

ACKNOWLEDGMENTS

The authors would like to acknowledge support from the National Science Foundation under Grant No. EFRI-1641078. We would also like to thank Ben Goldsberry for providing support and insights on the demonstration problem, Applied Research Labs for the use of their computational power, and both Clint Morris and Conner Sharpe for assisting the creative process with their openness to discussion of interesting problems. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

REFERENCES

[1] T. S. Chang, A. C. Ward, J. Lee and E. H. Jacox, "Conceptual Robustness in Simultaneous Engineering: An Extension of Taguchi's Parameter Design," *Research in Engineering Design,* vol. 6, no. 4, pp. 211-222, 1994.

[2] W. Chen and K. Lewis, "A Robust Design Approach for Achieving Flexibility in Multidisciplinary Design," *AIAA Journal,* vol. 37, no. 8, pp. 982-989, 1999.

[3] M. Kalsi, K. Hacker and K. Lewis, "A Comprehensive Robust Design Approach for Decision Trade-Offs in Complex Systems Design," *ASME Journal of Mechanical Design,* vol. 123, no. 1, pp. 1-10, 2001.

[4] H. Liu, W. Chen, M. J. Scott and K. Qureshi, "Determination of Ranged Sets of Design Specifications by Incorporating Design-Space Heterogeneity," *Engineering Optimization,* vol. 40, no. 11, pp. 1011-1029, 2008.

[5] J. H. Panchal, M. Gero Fernandez, C. J. J. Paredis, J. K. Allen and F. Mistree, "An Interval-Based Constraint Satisfaction (IBCS) Method for Decentralized, Colloborative Multifunctional Design," *Concurrent Engineering,* vol. 15, no. 3, pp. 309-323, 2007.

[6] H. Choi, D. L. McDowell, J. K. Allen, D. Rosen and F. Mistree, "An Inductive Design Exploration Method for Robust Multiscale Materials Design," *ASME Journal of Mechanical Design,* vol. 130, no. 3, p. 031402, 2008.

[7] D. Rosen, "A Set-Based Design Method for Material-Geometry Structures by Design Space Mapping," in *ASME IDETC Design Automation Conference*, Boston, MA, Paper Number: DETC2015-467602015.

[8] R. Arroyave, S. L. Gibbons, E. Galvan and R. Malak, "The Inverse Phase Stability Problem as a Constraint Satisfaction Problem: Application to Materials Design," *JOM,* vol. 68, no. 5, p. 13851395, 2016.

[9] P. Backlund, D. W. Shahan and C. C. Seepersad, "Classifier-guided Sampling for Discrete Variable, Discontinuous Design Space Exploration: Convergence and Computational Performance," *Engineering Optimization,* vol. 47, no. 5, pp. 579-600, 2015.

[10] D. W. Shahan and C. C. Seepersad, "Bayesian Network Classifiers for Set-Based Collaborative Design," *Journal of Mechanical Design,* vol. 134, no. 7, p. 071001, 2012.

[11] J. Matthews, T. Klatt, C. Morris, C. C. Seepersad and M. R. Haberman, "Hierarchical Design of Negative Stiffness Metamaterials Using a Bayesian Network Classifier," *Journal of Mechanical Design,* vol. 138, no. 4, p. 041404, 2016.

[12] C. Morris and C. C. Seepersad, "Identification of High Performance Regions of High-Dimensional Design Spaces with Materials Design Applications," in *IDETC Design Automation Conference*, Cleveland, OH, Paper Number: DETC2017-67769, 2017.

[13] C. Sharpe, C. Morris, B. Goldsberry, C. C. Seepersad and M. R. Haberman, "Bayesian Network Structure Optimization for Improved Design Space Mapping for Design Exploration with Materials Design Applications," in *IDETC Design Automation Conference*, Cleveland, OH, Paper Number: DETC2017-67643, 2017.

[14] Backlund, P., D.W. Shahan, and C.C. Seepersad, 2015, “Classifier-guided Sampling for Discrete Variable, Discontinuous Design Space Exploration: Convergence and Computational Performance, *Engineering Optimization,* Vol. 47, No. 5, pp. 579-600.

[15] D. W. Scott, Multivariate Density Estimation, New York: John Wiley & Sons Inc., 1992.

[16] A. Perez, P. Larranga and I. Inza, "Bayesian Classifiers Based on Kernel Density Estimation," *International Journal of Approximate Reasoning,* vol. 50, pp. 341-362, 2009.

[17] A. K. Jain and M. D. Ramaswami, "Classifier Design with Parzen Windows," in *Pattern Recognition and Artificial Intelligence*, North-Holland, Elsevier Science Publishers, 1988.

[18] J. S. Siminoff, Smoothing Methods in Statistics, New York: Springer-Verlag, 1996.

[19] K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe and W. Kegelmeyer, "Comparative Evaluation of Pattern

Recognition Techniques for Detection of Microcalcifications in Mammography," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 7, no. 6, pp. 1417-1436, 1993.

[20] M. Kubat, R. Holte and S. Matwin, "Machine Learning for the Dtection of Oil Spills in Satellite Radar Images," *Machine Learning,* vol. 30, no. 2-3, pp. 195-215, 1998.

[21] P. K. Chan, W. Fan, A. L. Prodromidis and S. J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems,* vol. 14, no. 6, pp. 67-74, 1999.

[22] N. Saliya, T. M. Khoshgoftaar and J. Van Hulse, "A Study on the Relationships of Classifier Performance Metrics," in *IEEE International Conference on Tools with Artificial Intelligence*, Newark, 2009.

[23] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[24] T. E. Fawcett and F. Provost, "Adaptive Fraud Detection," *Data Mining and Knowledge Discovery,* vol. 3, no. 1, pp. 291-316, 1997.

[25] N. Japkowicz, C. Myers and M. Gluck, "A Novelty Detection Approach to Classification," in *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada, 1995.

[26] C. X. Ling and C. Li, "Data Mining for Direct Marketing: Problems and Solutions," in *Dir*, International Conference on Knowledge Discovery and Data Mining, 1998.

[27] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume and C. Brunk, "Reducing Misclassification Costs," in *Proceedings of the Eleventh International Conference on Machine Learning*, New Brunswick, 1994.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[29] P. Domingos, "Metacost: A General Method for Making Classifiers Cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 1999.

[30] A. Estabrooks and N. Japkowicz, "A Mixture-of-Experts Framework for Concept-Learning from Imbalanced Data Sets," in *Proceedings of the 2001 Intelligent Data Analysis Conference*, Cascais, Portugal, 2001.

[31] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

[32] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Data Sets: One-Sided Sampling," in *Proceedings of the Fourteenth International Conference on Machine Learning*, Rome, 1997.

[33] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," *Intelligent Data Analysis,* vol. 6, pp. 429-449, 2002.

[34] H. Han, W. Y. Wang and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing ICIC 2005*, 2005.

[35] C. Bunkhumpornpat, K. Sinapiromsaran and L. Chidchanok, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009.

[36] E. Ramentol, Y. Caballero, R. Bello and F. Herrera, "SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory," *Knowledge and Information Systems,* vol. 33, no. 2, pp. 245-265, 2012.

[37] J. A. Saez, J. Luengo, J. Stefanowski and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences,* vol. 291, pp. 184-203, 2015.

[38] J. R. Willis, "The nonlocal influence of density variations in a composite," *International Journal of Solids and Structures Topics in Continuum Mechanics,* vol. 21, no. 7, pp. 805-817, 1985.

[39] J. A. Kong, "Theorems of bianisotropic media," *Proceedings of the IEEE,* vol. 60, no. 9, pp. 1036-1046, 1972.

[40] C. F. Sieck, A. Alù and M. R. Haberman, "Origins of Willis Coupling (Bianisotropy) in Acoustic Metamaterials through Source-Driven Homogenization," *Phys. Rev. B.,* vol. 96, no. 10, p. 104303, 2017.

[41] Shahan, D. and C. C. Seepersad, "Sequential Sampling with Kernel-Based Bayesian Network Classifiers," *ASME IDETC/CIE Design Automation Conference*, Washington, DC, Paper Number: DETC2011-48318, 2011.