# Signal Classification for Ground Penetrating Radar Using Sparse Kernel Feature Selection

Wenbin Shao, Abdesselam Bouzerdoum, *Senior Member, IEEE*, and Son Lam Phung, *Member, IEEE*

***Abstract*—This paper addresses the problem of feature selection for the classification of ground penetrating radar signals. We propose a new classification approach based on time–frequency analysis and sparse kernel feature selection. In the proposed approach, a time–frequency or a time-scale transform is first applied to the one-dimensional radar trace. Sparse kernel feature selection is then employed to extract an optimum set of features for classification. The sparse kernel method is formulated as an underdetermined linear system in a high-dimensional space, and the category labels of the training samples are used as measurements to select the most informative features. The proposed approach is evaluated through an industrial application of assessing railway ballast fouling conditions. Experimental results show that the proposed combination of sparse kernel feature selection and support vector machine classification yields very high classification rates using only a small number of features.**

***Index Terms*—Ground penetrating radar, pattern classification, sparse kernel feature selection.**

## I. Introduction

GROUND penetrating radar (GPR) is a geophysical testing tool used for nondestructive imaging of buried objects beneath the shallow earth surface and inside visually impenetrable structures [1]–[3]. It has been widely deployed in many application areas, such as archaeological explorations [4], [5], detection and monitoring of belowground biological structures [6], [7], glacier and ice sheet investigation [8], mineral resource evaluation [9], and land mine detection [10], [11]. GPR detects buried objects by transmitting electromagnetic waves, which radiate from the transmitting antenna and propagate into the subsurface. The wave is partially reflected back toward the receiving antenna when encountering an object whose electrical properties differ from that of the surrounding material. The reflected waves from buried objects form a nonstationary signal that captures electromagnetic characteristics of the objects.

In recent years, researchers have been actively investigating new applications [12]–[15], signal processing techniques [16]–[19], and hardware design [20], [21] for GPR. Among the processing techniques, time–frequency transforms have been widely used in radar and sonar signal processing [22], [23]. For example, Sun and Li proposed time–frequency localized features based on over-complete wavelet packet transform [24] and Savelyev *et al.* employed features extracted from Wigner–Ville transform [12] for land mine detection. Al-Qadi *et al*. considered the short-time Fourier transform (STFT) for ballast assessment [25]. They relied on visual inspection to compare the STFT images with the railtrack ground truth; they concluded that the STFT can be effective for ballast condition assessment. In [26], Lai and Poon presented two examples to analyze the variation in frequency content in STFT images with respect to material properties. Sinha *et al*. presented a method using the continuous wavelet transform to compute a time–frequency map for nonstationary seismic data [27].

In the past decade, compressed sensing (CS) has emerged as a powerful signal processing paradigm that allows a sparse signal to be exactly reconstructed from under-sampled information [28]–[30]. The CS theory is related to sparse representation (SR), which aims to find an efficient signal decomposition by expressing a signal as a linear combination of a few signal atoms chosen from an over-complete dictionary. Both CS and SR have attracted considerable interest from researchers in a wide range of areas, such as astronomical data compression [31], cognitive radar design [32], hyperspectral imaging [33], underwater sensor networks [34], and video streaming [35]. In [36], Wright *et al*. proposed an SR-based approach for face recognition. Their approach builds an over-complete dictionary with training samples and represents a test sample using a linear combination of the training samples from the same class. Ma and Le Dimet employed CS to deblur highly incomplete measurements in aerospace remote sensing [37]. In [38], Tang *et al*. proposed a two-stage approach for through-the-wall radar image-formation using CS. In [20] and [39], CS was applied to design stepped-frequency continuous-wave GPR systems.

In this paper, we propose a sparse kernel feature selection approach, based on CS and sparse signal representation, for GPR signal classification, where the time–frequency or time-scale representation is mapped into a high-dimensional feature space for feature selection. Our approach differs from the SR approach in [36] and [40] in a way the measurements are represented. In the SR approach, an unknown test sample is represented by a small fraction of training samples; classification is determined by the class that gives the minimal residual. By contrast, in the proposed approach, the category labels of the training samples are used as measurements in a

Manuscript received May 03, 2014; revised June 05, 2014; accepted July 13, 2014. Date of publication August 26, 2014; date of current version January 21, 2015. This work was supported by a grant from the Australian Research Council.

The authors are with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: ws909@uowmail.edu.au; a.bouzerdoum@ieee.org; s.phung@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2014.2341605

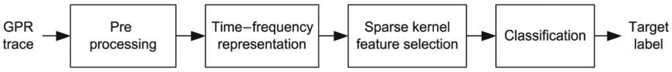

GPR trace
Pre processing
Time−frequency representation
Sparse kernel feature selection
Classification
Target label

Fig. 1. Block diagram of the proposed automatic classification system.

CS framework to select the salient features for classification. The selected features are then used as input to a support vector machine (SVM) classifier.

This paper is organized as follows. Section II explains the proposed sparse kernel feature selection approach. Section III applies the proposed approach to classification of GPR signals for railway ballast assessment. The proposed sparse kernel feature selection approach is assessed with different time–frequency and scale signal representations, three different kernels, and three classifiers. It is also compared with the local maxima feature selection method. Section IV gives concluding remarks.

## II. Proposed Approach

The proposed system, which is aimed at automatic classification of GPR signals, includes four main stages: 1) pre-processing; 2) time–frequency representation; 3) feature extraction; and 4) classification. The system block diagram is shown in Fig. 1. First, some basic signal processing techniques are applied to reduce the artifacts introduced by the GPR; they include DC component removal, resampling and time shifting. Next, the one-dimensional GPR traces are transformed into two-dimensional signals using a time–frequency or time-scale representation. In the third stage, CS is used to select the salient features to be used in the classification stage. In our approach, the feature selection is formulated as an underdetermined linear system in a high-dimensional feature space, and the salient features are found by solving an $\ell_0$ minimization problem. In the final stage, selected features are fed as inputs to a classifier, which classify the GPR traces into different types of ballast fouling. There are many pattern classifiers that can be used in this stage; however, this paper employs SVMs as the main classification tool for their excellent generalization ability.

In the following, we first introduce several time–frequency and time-scale techniques for signal representation, namely the short-time Fourier transform, S-method, and wavelet transform. The proposed sparse kernel feature selection method is presented in Section II-B, along with a local maxima approach. Also described in this section are two methods based on class separability for analyzing the effectiveness of the selected features. Finally, three classification methods are briefly introduced in Section II-C.

### A. Time–Frequency and Time–Scale Signal Representation

Nonstationary signals are usually analyzed using a time–frequency or time-scale representation. Here, we consider the short-time Fourier transform, S-method, and wavelet transform for feature extraction. Thus, a brief introduction to each method is warranted.

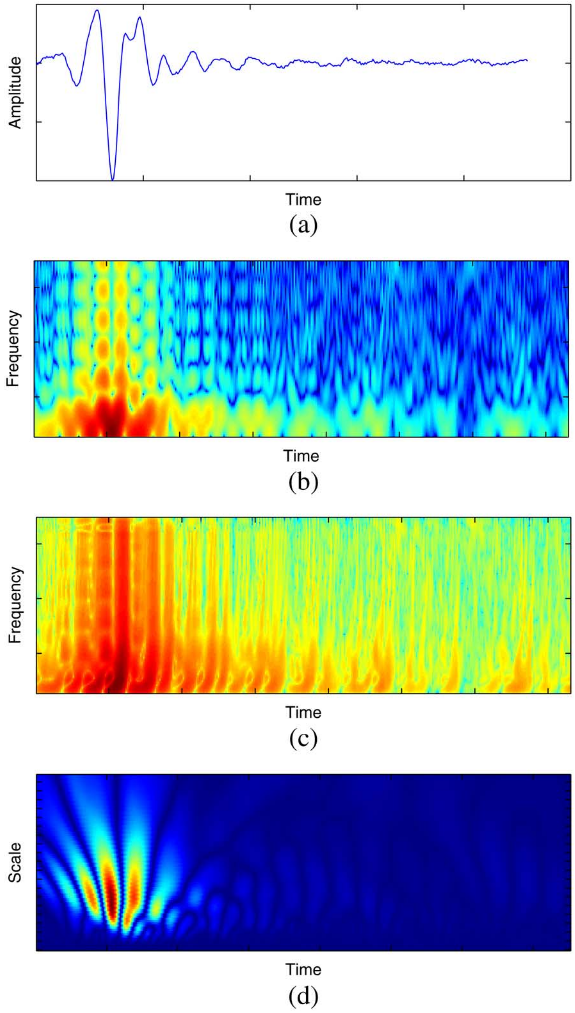


Fig. 2. Time–frequency and time-scale signal representations of a GPR signal. (a) A GPR trace. (b) STFT representation. (c) S-method representation. (d) Continuous wavelet transform representation.

*1) Short-Time Fourier Transform:* The STFT of a GPR trace $s(n)$ is defined as

$$\mathbf{S}(m,\omega) = \sum_{n=-\infty}^{\infty} s(n)h(n-m)e^{-j\omega n} \qquad (1)$$

where $h(n)$ is a window function of length $N$ and $\mathbf{S}(m,\omega)$ is a complex function containing the phase and magnitude information of the signal. The spectrogram of a radar trace is given by the squared magnitude of the STFT

$$S(m,\omega) = |\mathbf{S}(m,\omega)|^2 . \qquad (2)$$

In practice, (1) is implemented using the discrete Fourier transform (DFT), which leads to a discrete STFT in time and frequency $\mathbf{S}(m,k)$. Thus, the spectrogram can be represented as an image whose size is determined by the length of the trace and the size of the DFT. Fig. 2(b) illustrates the spectrogram of the GPR trace shown in Fig. 2(a).

There are many window types that can be used to smooth out the spectrogram, e.g., rectangular window, Hamming window, Hann window, Gaussian window, Chebyshev window, and Blackman window. This paper employs the Kaiser window because it provides a good trade-off between the mainlobe width and sidelobe height. The Kaiser window of length $N = 2N_h + 1$ is given by

$$h(n) = \frac{I_0(\beta\sqrt{1-[n/N_h]^2})}{I_0(\beta)}, \quad \text{for } n \le |N_h| \tag{3}$$

where $\beta$ is a parameter controlling the shape of the window and $I_0(\beta)$ is the modified Bessel function of the first kind and zeroth order.

*2) S-Method:* The S-method is a time–frequency representation derived based on smoothed pseudo Wigner distributions, in which the cross-terms are reduced or removed [41]. It combines the values of the STFT along the frequency axis. Mathematically, the discrete S-method is defined as

$$\mathbf{SM}(m,k) = \sum_{l=-N_h}^{N_h} h(l)\,\mathbf{S}(m,k+l)\,\mathbf{S}^*(m,k-l) \tag{4}$$

where $m$ and $k$ are the variables for time and frequency, $h(l)$ is a window function of length $N = 2N_h + 1$, and $^*$ denotes complex conjugate. The S-method has been applied in a number of areas, such as human gait classification [42], radar signal decomposition [43], and fault detection [44]. Fig. 2(c) shows the output of the S-method when applied to the GPR signal of Fig. 2(a).

*3) Wavelet Transform:* Wavelets provide a time-scale representation of signals. They are widely used for multiresolution signal analysis [45]. The continuous wavelet transform is defined as

$$W_\psi(a,\tau) = \int_{-\infty}^{\infty} s(t)\,\psi_{a,\tau}(t)dt \tag{5}$$

where $\psi(t)$ is the wavelet function given by

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \tag{6}$$

and $a$ and $\tau$ are the scale and translation parameters, respectively. Commonly used wavelets include Mexican hats, Haar, and Daubechies, to name a few. A time-scale representation of the GPR trace in Fig. 2(a), using the Daubechies wavelet, is shown in Fig. 2(d).

The discrete wavelet transform of signal $s(n)$ is given by

$$W_\varphi(j_0,k) = \frac{1}{\sqrt{N}}\sum_n s(n)\varphi_{j_0,k}(n) \tag{7}$$

$$W_\psi(j,k) = \frac{1}{\sqrt{N}}\sum_n s(n)\psi_{j,k}(n), \quad \text{for } j \ge j_0 \tag{8}$$

where $j_0$ and $j$ are the integer-scale parameters, $k$ is the integer translation parameter, $\varphi(n)$ is a real, square-integrable scaling function, and $\psi(n)$ is a wavelet function. Equation (7) defines the approximation coefficients and (8) specifies the detail coefficients.

### *B. Feature Extraction and Analysis*

This section describes feature extraction from a time–frequency or time-scale representation. We first present the proposed sparse kernel approach for feature selection. Then, we describe an existing approach called local maxima, which is used for comparison. Finally, we present two class separability criteria for evaluating the effectiveness of the selected features.

*1) Sparse Kernel Feature Selection:* In the proposed sparse kernel approach, feature selection is formulated as solving an underdetermined system of linear equations; the features are selected in a high-dimensional feature space. Consider $P$ training samples that belong to $C$ classes

$$\{(s_1,y_1),(s_2,y_2),\ldots,(s_P,y_P)\} \tag{9}$$

where $s_i$ is the $i$th GPR trace and $y_i \in \{1,2,\ldots,C\}$ is the class label. The corresponding time–frequency signals for the $P$ samples are denoted as

$$\{S_1,S_2,\ldots,S_P\}.$$

Each column of $S_i$ is an estimate of the frequency content of a time-localized section of the GPR trace $s_i$; time increases across the columns.

Before explaining the proposed approach, we first present a brief introduction to the traditional SR. SR expresses a signal as a linear combination of atoms chosen from an over-complete dictionary $\mathbf{X} \in \mathbb{R}^{P\times Q}$, with $P < Q$. The sparsity of a discrete-time signal $\mathbf{w} \in \mathbb{R}^Q$ is defined as the number of nonzero elements in $\mathbf{w}$. The $\ell_0$ pseudonorm, denoted as $\|\mathbf{w}\|_0$, is usually used as a measure of sparsity. If $\|\mathbf{w}\|_0 = q$, the vector $\mathbf{w}$ is called $q$-sparse. Suppose that the signal $\mathbf{y}$ is to be modeled with a small fraction of atoms from the dictionary $\mathbf{X}$, the process can be formulated as

$$\mathbf{y} = \mathbf{X}\mathbf{w}. \tag{10}$$

Since $P < Q$, (10) defines an underdetermined system of linear equations, and hence the recovery of $\mathbf{w}$ from this equation is ill-conditioned. However, a sparse vector $\mathbf{w}$ can be recovered by solving the following $\ell_0$ minimization problem:

$$\min \|\mathbf{w}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{X}\mathbf{w}. \tag{11}$$

This is a nondeterministic polynomial-time (NP) hard problem where an exhaustive search requires high-computational cost [30]. To solve this problem, several alternative approaches have been proposed, such as reweighted $\ell_1$ minimization [46], gradient projection [47], and orthogonal matching pursuit (OMP) [48]. In this paper, we adopt OMP algorithm to solve (11).

The key aspect now is how to adopt the SR paradigm for feature selection. Recall that our aim is to find the localized salient frequencies from the training data. This is equivalent to finding a representative subset of frequencies, which best describes the relationship between the frequency components and the class labels. Therefore, we construct the dictionary matrix $\mathbf{X}$ using time–frequency or time-scale representations of GPR signals, and use the class label set as the signal $\mathbf{y}$. Each row of the measurement matrix $\mathbf{X}$ contains the time–frequency or time-scale representation of a trace. That is, the $i$th row of $\mathbf{X}$ is the vector form of $S_i$, which is obtained by a lexicographical ordering. Thus, each column of $\mathbf{X}$ represents one point in the time–frequency or time-scale representation, i.e., one frequency or one scale at one particular time instant. The vector $\mathbf{y}$ is not a signal in the traditional sense of SR, but a collection of class labels from the training data, $y_i \in \{1, 2, \ldots, C\}$.

Let $\Phi$ be a mapping function that projects the features in the input space to a high-dimensional feature space. The sparse kernel feature selection is formulated as

$$\min \|\tilde{\mathbf{w}}\|_0 \quad \text{subject to} \quad \Phi(\mathbf{y}) = \Phi(\mathbf{X})\tilde{\mathbf{w}}. \tag{12}$$

We focus on the positions of nonzero coefficients in $\tilde{\mathbf{w}}$, not the coefficient values, because the coefficient positions indicate the selected features.

The difficulty in solving (12) using the OMP is to compute the mapping $\Phi$ from input space to feature space. The feature space is high dimensional, which results in an unaffordable computational cost. OMP can be expressed in dot product form. Here, we employ the kernel trick to evaluate the dot products in the feature space, without having to compute the mapping explicitly; that is

$$\kappa(\mathbf{z}, \mathbf{z}') = \langle \Phi(\mathbf{z}), \Phi(\mathbf{z}') \rangle \tag{13}$$

where $\kappa$ represents a positive semidefinite kernel. There are several commonly used kernels, such as linear kernel, polynomial kernel, and radial basis function (RBF) kernel [49]–[51].

1) Linear kernel
$$\kappa(\mathbf{z}, \mathbf{z}') = \mathbf{z}^T \mathbf{z}'. \tag{14}$$

2) Polynomial kernel
$$\kappa(\mathbf{z}, \mathbf{z}') = (\mathbf{z}^T \mathbf{z}' + 1)^p \tag{15}$$
where $p$ is the polynomial degree, $p \in \mathbb{N}$.

3) RBF kernel
$$\kappa(\mathbf{z}, \mathbf{z}') = e^{-\gamma \|\mathbf{z}-\mathbf{z}'\|^2} \tag{16}$$
where $\gamma$ is a positive scalar that controls the kernel bandwidth.

In practice, the kernel is chosen empirically according to prior knowledge of the data and the application [50]. The number of data samples and features also affects the choice of the kernel.

The main steps of the OMP algorithm using the kernel trick are explained as below.

1) Initialization
   a) the iteration index ($j = 1$);
   b) a Gram matrix $\mathbf{G}$ with respect to $\mathbf{x}_i$, where $\mathbf{x}_i$ denotes the $i$th column of $\mathbf{X}$, $i = 1, 2, \ldots, Q$;
   c) a column vector $\hat{\mathbf{r}}$ whose $i$th element is calculated by $\hat{r}_i = \kappa(\mathbf{x}_i, \mathbf{y})$;
   d) an empty vector $\boldsymbol{\lambda}_0 = \emptyset$; and
   e) the solution $\mathbf{x}_0 = 0$.
2) For the $j$th iteration, locate the atom $\Phi(\mathbf{x}_j)$ ($\mathbf{x}_j$ is a column of $\mathbf{X}$) that satisfies
$$i^* = \arg\max_i \left( \left| \hat{r}_i - \mathbf{w}_{j-1}^\top \mathbf{G}_{[\boldsymbol{\lambda}_{j-1},\, i]} \right| \right) \tag{17}$$
where $\mathbf{G}_{[\boldsymbol{\lambda}_{j-1},\, i]}$ represents the elements of matrix $\mathbf{G}$ in rows $\boldsymbol{\lambda}_{j-1}$ of column $i$. Append $i^*$ to the list of previously selected atom indices, $\boldsymbol{\lambda}_j = [\boldsymbol{\lambda}_{j-1},\, i^*]$.
3) Obtain the solution $\mathbf{w}_j = \mathbf{G}^{-1}_{[\boldsymbol{\lambda}_j,\, \boldsymbol{\lambda}_j]} \hat{\mathbf{r}}_{[\boldsymbol{\lambda}_j]}$.
4) Increase $j$ by 1 and repeat Steps 1)–3) until the predefined sparsity for $\mathbf{w}$ is reached.

The above algorithm is summarized in Fig. 3.

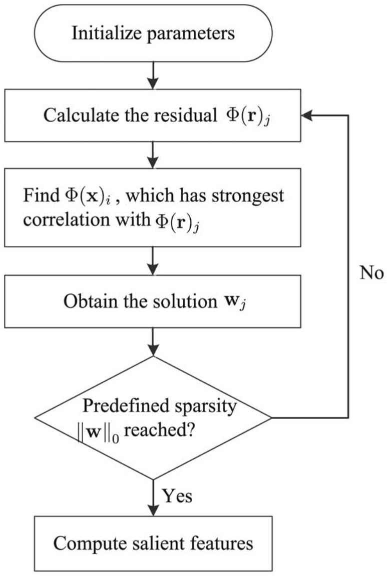

Initialize parameters
Calculate the residual $\Phi(\mathbf{r})_j$
Find $\Phi(\mathbf{x})_i$, which has strongest correlation with $\Phi(\mathbf{r})_j$
Obtain the solution $\mathbf{w}_j$
Predefined sparsity $\|\mathbf{w}\|_0$ reached?
No
Yes
Compute salient features

Fig. 3. The main steps of the proposed feature selection algorithm based on OMP and kernel trick.

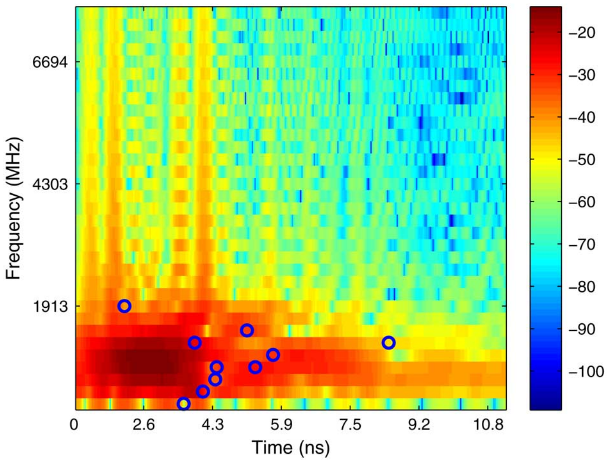


Fig. 4. Examples of salient features selected by the sparse kernel. Each circle represents a feature point.

For a time–frequency representation, at each iteration OMP identifies the most significant feature (i.e., a point in the time–frequency space representing one frequency at one instant of time) that leads to the best approximation of the target signal; it ensures that the same feature is not selected twice. Selected features are ranked in the same order as the OMP iterations. The time information is implicitly included in the indices of the nonzero coefficients of $\tilde{\mathbf{w}}$. The same principle applies for time-scale representations. Fig. 4 shows an example of features selected using the sparse kernel approach from the STFT spectrogram.

*2) Local Maxima:* Here, the selected features using the proposed sparse kernel representation are compared with features extracted at the local maxima of the time–frequency or time-scale representation. The local maxima approach extracts features using morphological dilation. Let $a$ be a structuring element, which is a binary image of any arbitrary shape and

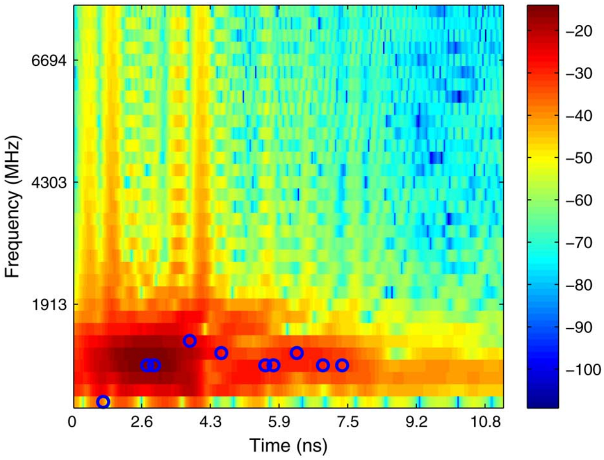


Fig. 5. Examples of salient features selected by the local maxima. Each circle represents a feature point.

size. The dilation of an image, e.g., the spectrogram, $S$ by $a$, denoted by $S \oplus a$, is defined as

$$[S \oplus a](m,n) = \max\{S(m-p,n-q)+a(p,q) | (m-p,n-q) \in \mathcal{D}_S; (p,q) \in \mathcal{D}_a\}, \quad (18)$$

where $\mathcal{D}_S$ and $\mathcal{D}_a$ are the domains of $S$ and $a$, respectively. In this approach, local maxima are considered as the salient features. When the same local maximum is shared by several traces with different magnitudes, the average magnitude is used to sort the local maxima. Clustering is applied to reduce the number of local maxima. Magnitude features are then extracted at the local maxima. Because each local maximum corresponds to a salient frequency, the magnitudes are considered as the strengths of the features.

Note that a local maxima approach was proposed in [14] to extract features from the 1-D Fourier transform. In this paper, we extend this approach to the 2-D time–frequency and time-scale representations. Fig. 5 shows an example of features selected using the local maxima approach from the STFT spectrogram.

*3) Class Separability:* To assess the effectiveness of the sparse kernel features, we investigate two class separability criteria. The first criterion is based on multiple discriminant analysis (MDA) [52]. Suppose that we have $C$ sets of labeled features

$$\mathcal{D}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{N_i}^i\}, \quad i \in \{1,2,\ldots,C\} \quad (19)$$

where each set corresponds to one class label. For class $i$, the centroid is given by

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{x}_j^i \quad (20)$$

and the data scatter is defined as

$$\tilde{\mathbf{\Sigma}}_i = \sum_{j=1}^{N_i}(\mathbf{x}_j^i - \mu_i)(\mathbf{x}_j^i - \mu_i)^\mathsf{T}. \quad (21)$$

The within-class and between-class scatter $\mathbf{\Sigma}_\mathrm{w}$ and $\mathbf{\Sigma}_\mathrm{b}$, respectively, are given by

$$\mathbf{\Sigma}_\mathrm{w} = \sum_{i=1}^{C}\tilde{\mathbf{\Sigma}}_i \quad (22)$$

and

$$\mathbf{\Sigma}_\mathrm{b} = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^\mathsf{T} \quad (23)$$

where $\mu$ is the global mean of all data sets. The objective of MDA is to maximize the ratio of between-class scatter to the within-class scatter. We use the determinant of the scatter matrix as a measure of the scatter, and define a scalar score function $J(\mathbf{w})$ as follows:

$$J(\mathbf{w}) = \frac{|\mathbf{w}^\mathrm{T}\mathbf{\Sigma}_\mathrm{b}\mathbf{w}|}{|\mathbf{w}^\mathrm{T}\mathbf{\Sigma}_\mathrm{w}\mathbf{w}|} \quad (24)$$

where $\mathbf{w}$ satisfies

$$\mathbf{\Sigma}_\mathrm{b}\mathbf{w} = \lambda\mathbf{\Sigma}_\mathrm{w}\mathbf{w}. \quad (25)$$

We also use mutual information to measure the class separability. Mutual information between two variables is the reduction in the uncertainty of one variable given that the other variable is known. Let $x$ be the observation (the feature) and $y$ be the class label. The mutual information between the feature and the class label is defined as

$$\mathrm{MI}(y,x) = \sum_y\sum_x \mathcal{P}(y,x)\log_2\frac{\mathcal{P}(y,x)}{\mathcal{P}(y)\mathcal{P}(x)} \quad (26)$$

where $\mathcal{P}(y,x)$ is the joint probability density function, and $\mathcal{P}(x)$ and $\mathcal{P}(y)$ are probability mass functions.

### C. Classification

There exist many pattern classifiers, including linear discriminant analysis, $K$-nearest neighbors ($K$-NNs), Bayes classifier, neural networks, and SVMs [52]. In this paper, the SVMs are used as the main classification tool. The results are compared with those of $K$-NN and a linear sparse classifier (LSC).

*1) Support Vector Machine:* SVMs are originally formulated for two-class classification problems. Consider $P$ training samples from two classes: 1) $\{(x_1,y_1),(x_2,y_2),\ldots,(x_P,y_P)\}$; 2) $y_i \in \{-1,+1\}$. If the classes are linearly separable in the input space, the decision function is written as

$$f(x) = \mathrm{sgn}(\langle\mathbf{w}, x_i\rangle + b) \quad (27)$$

where $\mathbf{w}$ is the vector normal to the hyperplane and $b$ is a bias term. In SVMs, the decision boundary is obtained from the training data by finding a separating hyperplane (represented by $\mathbf{w}$ and $b$) that maximizes the margins between the two classes.

The margin perpendicular to the hyperplane can be expressed as $2/\|\mathbf{w}\|$. Consequently, the problem is equivalent to minimizing

$$Q(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \quad (28)$$

subject to

$$y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1, \quad \text{for } i = 1, \ldots, P. \tag{29}$$

If the classes are not linearly separable, it is necessary to introduce nonnegative slack variables $\xi_i$ into the constraint in (29). This learning strategy is shown to increase the generalization capability of the classifier. SVMs can be applied to complex nonlinearly separable problems by projecting the data onto a high-dimensional space using kernel methods. For a detailed description of the SVM classifier, the reader is referred to [49] and [51].

*2) $K$-Nearest Neighbor:* The $K$-NN classifies a new sample by using the labels of the closest $K$ training samples [53]. The steps to classify a test sample $x$ are as follows:

1) specify $K$, the number of nearest neighbors;
2) calculate the distance, usually the Euclidean distance, between the test sample $x$ and all the other samples in the training set;
3) collect the class labels of the $K$-NNs;
4) assign the most frequent class label in the $K$-NNs to $x$.

The $K$-NN classifier is selected for comparison because it is a nonparametric algorithm that can achieve a high-classification accuracy when there are sufficient training samples. In pattern classification, the $K$-NN is often used as a baseline for comparison [52].

*3) Linear Sparse Classifier:* The LSC is based on the solution of (11). Equation (10) shows that the class label is represented by a linear combination of a small fraction of the coefficients from time–frequency or time-scale representations. When the feature selection is formulated directly in the input space, the sparse solution $\mathbf{w}$ can be used as a weight vector of an LSC classifier. Given a new test sample, the predicted class is given by

$$y_p = \langle \mathbf{x}, \mathbf{w} \rangle \tag{30}$$

where $y_p$ is the prediction score and $\mathbf{x}$ is a vector containing the time–frequency or time-scale representation of a trace.

## III. Experiments and Analysis

In this section, the proposed sparse kernel feature selection approach is evaluated on a GPR data set, which was collected for automatic railway ballast assessment. Section III-A describes the data set and the experimental methods. Section III-B analyzes the sparse kernel feature selection approach with the three time–frequency and time-scale signal representations discussed in Section II-A, different kernels, and different classifiers. Section III-C compares the proposed approach with the local maxima feature extraction approaches.

### A. Experimental Methods

The experiments were conducted on real data collected as part of a project for ballast fouling assessment at Wollongong, NSW, Australia. This project investigated new, noninvasive methods for assessing railway ballast conditions using GPR. The data were acquired on a track that was laid parallel to several existing tracks in service. Considering the time and cost, three of the most common ballast fouling conditions were investigated: 1) 50% clay fouling; 2) clean; and 3) 50% coal fouling. Here, the fouling material was measured using relative ballast fouling ratio. Three railtrack sections were constructed, and each was filled with the ballast of one fouling condition. Each section spanned a length of 2.0 m and a depth of 0.55 m; the width was equivalent to the existing ballast width.

TABLE I
Summary of the Wollongong Railway GPR Data Set

| Data subset | AH-200 mm | AH-300 mm | AH-400 mm |
|---|---|---|---|
| Condition | Dry | Dry | Wet |
| Antenna height | 200 mm | 300 mm | 400 mm |
| Clay ballast | 469 | 470 | 745 |
| Clean ballast | 477 | 478 | 642 |
| Coal ballast | 436 | 438 | 705 |
| Total traces | 1382 | 1386 | 2092 |

The antenna frequency is 800 MHz.

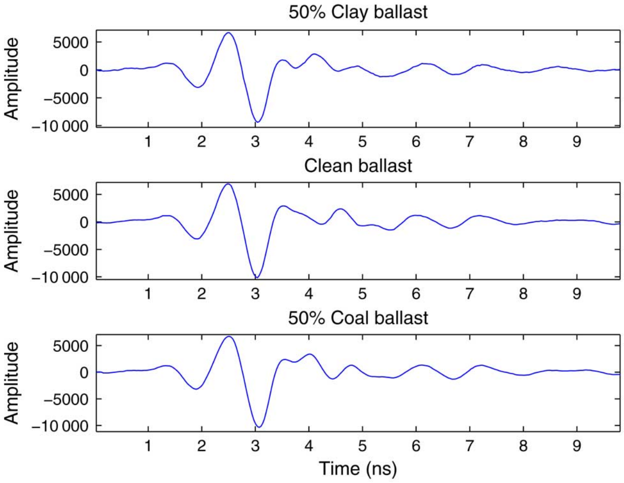


Fig. 6. Three traces from the Wollongong railway data set: from top to bottom, 50% clay ballast, clean ballast, and 50% coal ballast, respectively.

The antenna center frequency was set to 800 MHz. In preliminary railtrack surveys, this frequency was found to produce clearer GPR signals for ballast assessment than 1.2 GHz frequency. Three subsets of data were collected by varying the antenna height with respect to the ground: AH-200 mm data subset, AH-300 mm data subset, and AH-400 mm data subset, where AH stands for antenna height. The AH-200 mm and AH-300 mm data subsets were collected under dry ground condition, i.e., sunny weather and dry materials. The AH-400 mm data subset was acquired under wet condition, i.e., cloudy weather and water-saturated materials. A summary of the Wollongong railway data set is presented in Table I. The time-domain waveforms of three traces from different fouling types are shown in Fig. 6.

In our experiments, the performance is assessed using the classification rate as a function of the number of features. The classification rate is the percentage of test samples that are correctly classified

$$\text{CR} = \frac{N_c}{N_t} \times 100\% \tag{31}$$

where $N_t$ is the total number of test samples and $N_c$ is the number of correctly classified test samples. There are several ways to estimate the generalization ability of a classifier,

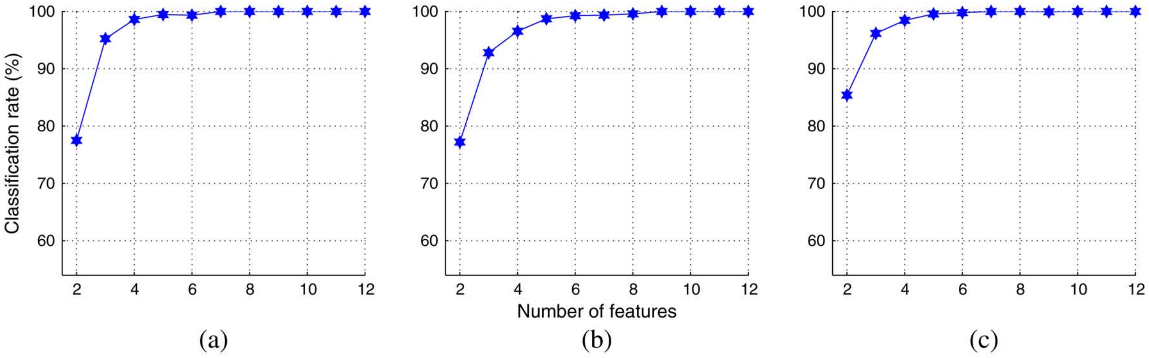


Fig. 7. Classification rates of features extracted using the short-time Fourier transform: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm data subsets.

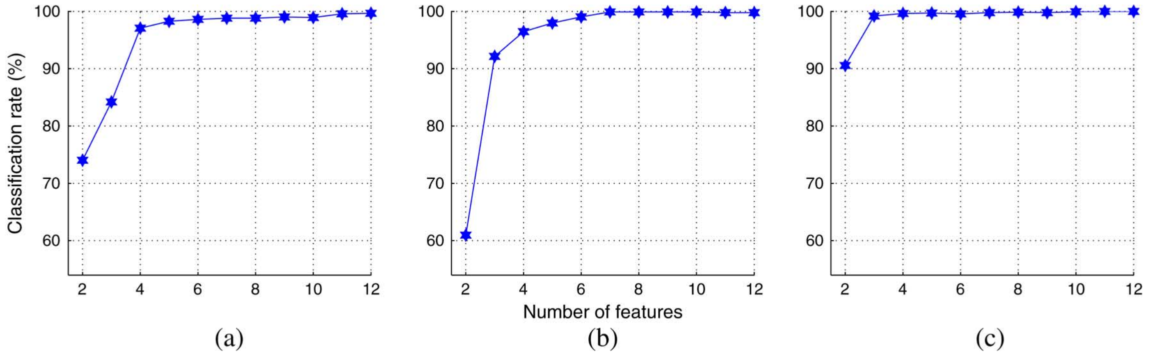


Fig. 8. Classification rates of features extracted using the S-method: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm data subsets.

such as cross validation, Vapnik–Chervonenkis dimension, and leave-one-out error rate estimators [51]. We employed fivefold cross-validation, considering its computational requirements and estimation reliability [54]. The data set is randomly divided into five partitions of approximately equal size. Four partitions are used for training and validation, and the remaining partition is used for testing the classifier. This step is repeated five times, each time using a different partition for testing. This procedure ensures that each sample is tested only once. The final classification rate is computed using the aggregate number of correctly classified samples across all the folds.

### *B. Analysis of Sparse Kernel Feature Selection*

This section investigates the performance of the proposed sparse kernel feature selection approach with different signal representations, different kernels, and different classifiers. In the following experiments, except in Section III-B2, the linear kernel is used in the sparse kernel approach feature selection.

*1) Different Signal Representations:* In this section, we analyze classification performance when the proposed sparse kernel method is applied to three different time–frequency and time-scale representations, namely the STFT, S-method, and discrete wavelet transform.

For the STFT, the Kaiser window parameter $\beta$ was set to 2. The results in Fig. 7 show that the classification rates reach $90.0\%$ with 3 or more features, and $98.0\%$ with 5 or more features on all three data subsets.

The classification performance using the S-method is shown in Fig. 8. Using five or more salient features leads to a classification rate of $98\%$ on all data subsets. When three or fewer features are used, the S-method achieves a lower classification rate than STFT on the AH-200 mm and AH-300 mm subsets, but obtains a higher rate on the AH-400 mm subset.

For the wavelet transform, different wavelets, including one Coiflets, three Daubechies, and two Symlets, are evaluated; these wavelets are chosen because of their shape similarity with GPR traces. The classification rates for different wavelets are presented in Fig. 9. On the AH-200 mm subset, the classification rates with the six wavelets differ widely when three or fewer features are used. However, the difference between the six wavelets decreases rapidly as the number of features increases; with seven or more features, the classification rates become almost identical for all six wavelet functions. Similar observations can be made on the AH-300 mm and AH-400 mm subsets. All the wavelets reach similar classification rates when five or more features are utilized. Among the six wavelets, the Daubechies wavelet of fifth order achieves the best performance on the three data subsets.

Table II summarizes the classification rates for the STFT, S-method, and wavelets features selected by the proposed sparse kernel method. The classification rate stabilizes when five or more features are used regardless of the time–frequency or time-scale representation. In general, the effects of different signal representations on the classification rate diminish as the number of features increases.

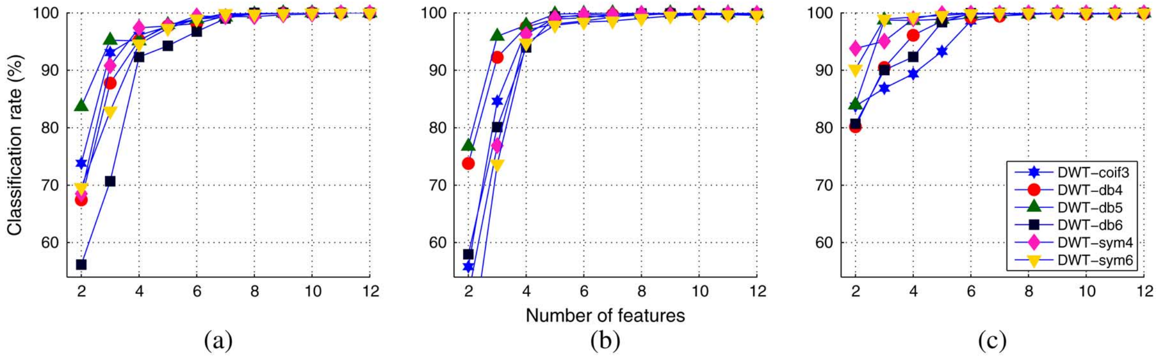


Fig. 9. Classification rates of features extracted using different wavelets: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm data subsets.

TABLE II

Classification Rates (%) for Different Numbers of Features on the AH-200 mm, AH-300 mm, and AH-400 mm Data Subsets

| Features | Number of features | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|---|
| STFT | AH-200 mm | $77.5 \pm 2.7$ | $98.6 \pm 0.8$ | $99.4 \pm 0.5$ | $\mathbf{100.0} \pm 0.0$ | $\mathbf{100.0} \pm 0.0$ |
| | AH-300 mm | $77.2 \pm 2.7$ | $96.6 \pm 1.2$ | $99.3 \pm 0.6$ | $99.6 \pm 0.4$ | $\mathbf{100.0} \pm 0.0$ |
| | AH-400 mm | $85.4 \pm 1.8$ | $98.5 \pm 0.7$ | $99.8 \pm 0.3$ | $\mathbf{100.0} \pm 0.0$ | $\mathbf{100.0} \pm 0.0$ |
| S-method | AH-200 mm | $74.0 \pm 2.8$ | $97.1 \pm 1.1$ | $98.6 \pm 0.8$ | $98.9 \pm 0.7$ | $98.9 \pm 0.7$ |
| | AH-300 mm | $61.0 \pm 3.1$ | $96.5 \pm 1.2$ | $99.0 \pm 0.6$ | $99.9 \pm 0.2$ | $99.9 \pm 0.2$ |
| | AH-400 mm | $\mathbf{90.1} \pm 1.6$ | $\mathbf{99.6} \pm 0.3$ | $99.6 \pm 0.4$ | $99.9 \pm 0.2$ | $99.9 \pm 0.1$ |
| DWT | AH-200 mm | $83.7 \pm 2.4$ | $95.1 \pm 1.4$ | $98.9 \pm 0.7$ | $99.8 \pm 0.3$ | $\mathbf{100.0} \pm 0.0$ |
| | AH-300 mm | $76.8 \pm 2.7$ | $97.9 \pm 0.9$ | $\mathbf{99.9} \pm 0.2$ | $99.9 \pm 0.2$ | $99.9 \pm 0.2$ |
| | AH-400 mm | $84.0 \pm 2.0$ | $98.7 \pm 0.6$ | $99.8 \pm 0.3$ | $99.8 \pm 0.3$ | $\mathbf{100.0} \pm 0.0$ |

The $95\%$ confidence interval is calculated using Student's $t$-distribution.

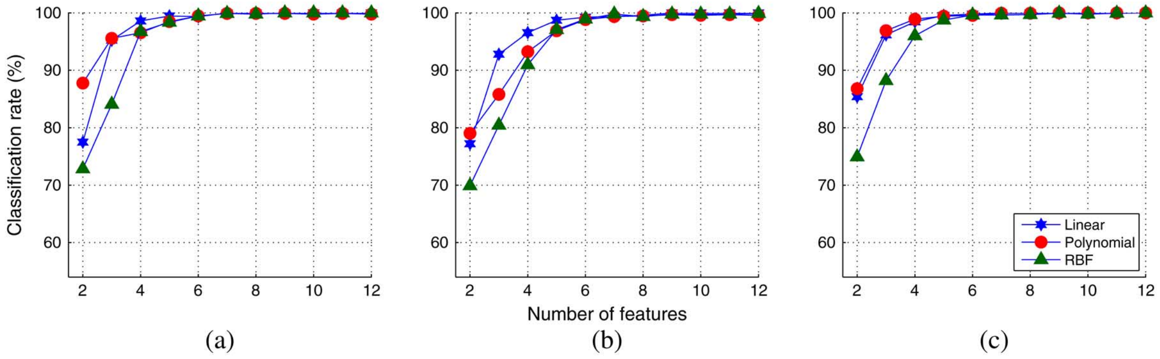


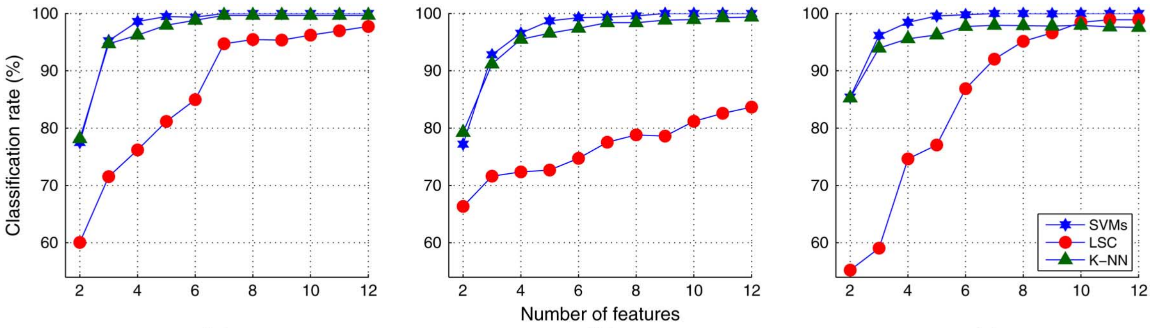
Fig. 10. Classification rates of linear, polynomial, and RBF kernels: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm data subsets.

*2) Different Kernels:* In this section, we analyze the effect of different kernels used in the proposed sparse kernel approach, including the linear kernel, polynomial kernel, and RBF kernel. Note that the definitions of the kernels are given in Section II-B.

Fig. 10 shows the classification rates when different kernels are used in the sparse kernel feature selection. On the AH-200 mm data subset, all kernels reach a classification rate of $96\%$ with four features. On the AH-300 mm subset, five features are required to achieve a CR of $96\%$. On the AH-400 mm subset, the linear kernel and the polynomial kernel yield very close classification rates, and they slightly outperform the RBF kernel. However, for all data subsets, all kernels achieve the same performance with six or more features.

*3) Different Classifiers:* In this section, the SVM classifiers are compared with two other approaches: the LSC and the $K$-NN classifier. The features are extracted using STFT and sparse kernel feature selection. During the experiments, the parameter $K$ of the $K$-NN is varied from 1 to 17 with a step of 2, whereas the SVM parameters are selected through a fivefold cross-validation. Fig. 11 shows the classification rates of the different classifiers as a function of the number of features. In general, the SVM and the $K$-NN classifiers outperform the LSC. However, the classification rate of the latter approaches those of the SVM and $K$-NN as the number of input features increases beyond 10. On the AH-300 subset, the performance of the LSC improves significantly and approaches those of SVM and $K$-NN when the number of features is increased to more than 40.

Table III presents the classification rates of the three classifiers for a select number of features. These results show that the SVM classifier achieves better classification performance than the other two classifiers. With only three features, the SVM classifier achieves a classification rate of $94.9\%$ across



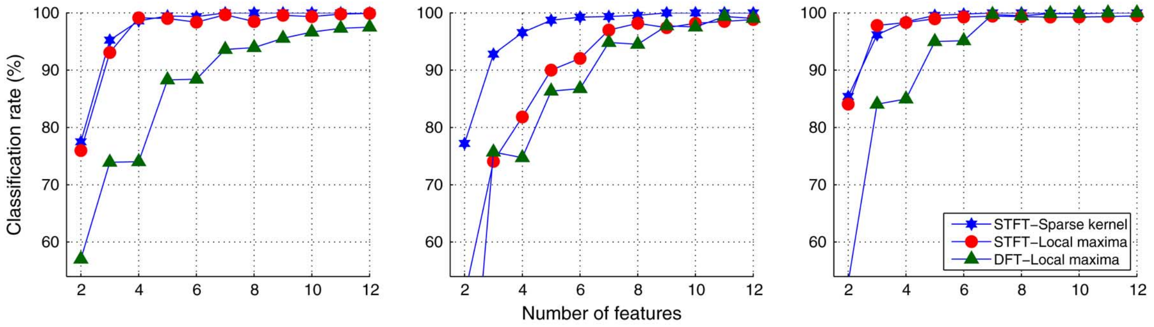
(a) (b) (c)

Fig. 11. Classification rates of SVM, LSC, and $K$-NN classifiers: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm data subsets.

TABLE III
CLASSIFICATION RATES OF SVM, $K$-NN, AND LSC USING STFT FEATURES

| Number of features | | 3 | 5 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|
| AH-200 mm | SVM | $95.2 \pm 1.4$ | $99.5 \pm 0.5$ | $100.0 \pm 0.0$ | $100.0 \pm 0.0$ | $100.0 \pm 0.0$ |
| | $K$-NN | $94.7 \pm 1.5$ | $97.9 \pm 0.9$ | $99.7 \pm 0.4$ | $99.7 \pm 0.4$ | $99.7 \pm 0.4$ |
| | LSC | $71.5 \pm 2.9$ | $81.2 \pm 2.5$ | $94.7 \pm 1.4$ | $95.3 \pm 1.4$ | $97.0 \pm 1.1$ |
| AH-300 mm | SVM | $92.8 \pm 1.7$ | $98.7 \pm 0.7$ | $99.4 \pm 0.5$ | $100.0 \pm 0.0$ | $100.0 \pm 0.0$ |
| | $K$-NN | $91.2 \pm 1.8$ | $96.6 \pm 1.2$ | $98.4 \pm 0.8$ | $98.8 \pm 0.7$ | $99.3 \pm 0.6$ |
| | LSC | $71.6 \pm 2.9$ | $72.7 \pm 2.9$ | $77.5 \pm 2.7$ | $78.6 \pm 2.6$ | $82.6 \pm 2.4$ |
| AH-400 mm | SVM | $96.2 \pm 1.0$ | $99.6 \pm 0.4$ | $100.0 \pm 0.0$ | $99.9 \pm 0.1$ | $100.0 \pm 0.0$ |
| | $K$-NN | $94.0 \pm 1.3$ | $96.2 \pm 1.0$ | $97.9 \pm 0.8$ | $97.8 \pm 0.8$ | $97.6 \pm 0.8$ |
| | LSC | $59.1 \pm 2.6$ | $77.1 \pm 2.2$ | $92.0 \pm 1.4$ | $96.6 \pm 1.0$ | $98.9 \pm 0.6$ |
| Combined set | SVM | $94.9 \pm 0.8$ | $99.3 \pm 0.3$ | $99.5 \pm 0.2$ | $100.0 \pm 0.1$ | $100.0 \pm 0.0$ |
| | $K$-NN | $93.4 \pm 0.9$ | $96.8 \pm 0.6$ | $97.9 \pm 0.5$ | $98.6 \pm 0.4$ | $98.7 \pm 0.4$ |
| | LSC | $66.3 \pm 1.6$ | $77.0 \pm 1.5$ | $82.8 \pm 1.3$ | $91.0 \pm 1.0$ | $93.6 \pm 0.8$ |

The 95% confidence interval is calculated using Student's $t$-distribution.



(a) (b) (c)

Fig. 12. Classification rates of different feature types (DFT local maxima, STFT local maxima, and sparse kernel features) on three data subsets: (a) AH-200 mm; (b) AH-300 mm; and (c) AH-400 mm.

all three data subsets, while the $k$-NN and the LSC yield classification rates of 93.4% and 66.3%, respectively. When seven features are used, the SVM, $k$-NN, and LSC obtain classification rates of 99.5%, 97.9%, and 82.8%, respectively. With 11 features, the classification rates of SVM, $k$-NN, and LSC on the combined data set reach 100.0%, 98.7%, and 93.6%, respectively.

### C. Comparison With Local Maxima Features

This section presents a comparison between the sparse kernel features and the local maxima features on the railway trace classification. First, the two types of features are assessed in terms of their classification rates. Then, the sparse kernel and local maxima feature extraction methods are evaluated in terms of their effectiveness using two class separability measures.

*1) Classification Rates Versus Number of Features:* In this experiment, the sparse kernel features are compared, in terms of classification accuracy, with 1-D and 2-D local maxima features. Both the sparse kernel and 2-D local maxima features are extracted from the STFT spectrograms, whereas the 1-D local maxima are extracted from the DFT magnitude spectrum. Fig. 12 illustrates the classification rates of the three feature types as a function of the number of features. The results indicate that on all three data sets, the classification accuracy increases as more features are added. When ten or more features are used, all approaches achieve classification rates above 97.0%. However, for the same number of features, the proposed sparse kernel feature selection method yields

TABLE IV
RATIO (IN dB) OF THE MDA CLASS SEPARABILITY SCORES OF SPARSE KERNEL FEATURES ($J_{\mathrm{SC}}$) AND STFT LOCAL MAXIMA FEATURES ($J_{\mathrm{LM}}$) WHEN FIRST $n$ FEATURES ARE USED

| First $n$ features | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $J_{\mathrm{SC}}/J_{\mathrm{LM}}$ | 14.6 | 47.0 | 39.0 | 75.1 | 95.7 | 106.2 |

TABLE V
MUTUAL INFORMATION BETWEEN THE EXTRACTED FEATURES AND THE LABELS

| Approach \ Feature | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Sparse kernel | 0.749 | 0.288 | 0.331 | 0.449 | 0.532 |
| STFT local maxima | 0.288 | 0.135 | 0.231 | 0.121 | 0.191 |

higher classification rates than the two local maxima methods. Furthermore, the performance of the proposed method is more stable across the three data subsets; that is, for the same number of features, the proposed method achieves similar classification rates on all three subsets.

*2) Analysis of Class Separability:* The effectiveness of the extracted features is evaluated by their ability to separate different classes. The class separability of the sparse kernel and STFT local maxima features is compared using two measures: the MDA-based separability score given in (24) and mutual information. Table IV presents the ratio (in dB) of the class separability scores between sparse kernel features and STFT local maxima features. The results indicate that the sparse kernel features have better class separability. Table V shows the mutual information between the features and the labels. For each of the first five features, the sparse kernel-based approach selects a feature with higher mutual information with the class label.

The STFT local maxima method selects feature points that are concentrated on the high magnitude area. By contrast, the sparse kernel approach utilizes correlations between class labels and training data; the extracted features points are not necessarily the local maxima.

## IV. CONCLUSION

This article presented a new sparse kernel feature selection approach for GPR signal classification using time–frequency or time-scale signal representation. Three types of signal representations were investigated for feature extraction, namely the STFT, S-method, and the wavelet transform. The feature selection problem was formulated as an SR of class labels in high-dimensional space. Kernel OMP was then employed to solve the SR problem. The proposed method was applied to GPR signal classification for assessing railway ballast fouling conditions. The experiments were conducted using three subsets of real GPR data acquired at different heights, under dry and wet conditions. Three classifiers were tested with the sparse kernel feature selection: they include SVM, $K$-NN, and LSC. The experimental results showed that both SVM and $K$-NN can achieve high classification rates with a small number of features and that the performance of all classifiers improves steadily as the number of selected features increases. However, the SVM classifier achieved consistently higher classification rates than the other two classifiers. Finally, when compared with the local maxima feature selection technique, the proposed sparse kernel feature selection approach was found to yield higher classification rates, more stable performance, and a higher degree of class separability.

## ACKNOWLEDGMENT

The Wollongong railway GPR data were collected as part of the Rail CRC-AT5 project, sponsored by CRC for Rail Innovation.

## REFERENCES

[1] A. P. Annan, "GPR—History, trends, and future developments," *Subsurf. Sens. Technol. Appl.*, vol. 3, no. 4, pp. 253–270, 2002.

[2] D. J. Daniels, D. J. Gunton, and H. F. Scott, "Introduction to subsurface radar," *IEE Proc. F Radar Signal Process.*, vol. 135, no. 4, pp. 278–320, 1988.

[3] H. M. Jol, Ed., *Ground Penetrating Radar Theory and Applications*, 1st ed. Amsterdam, The Netherlands: Elsevier, 2009.

[4] E. Pettinelli, P. M. Barone, E. Mattei, and S. E. Lauro, "Radio wave techniques for non-destructive archaeological investigations," *Contemp. Phys.*, vol. 52, no. 2, pp. 121–130, 2011.

[5] E. Pettinelli, P. M. Barone, E. Mattei, A. D. Matteo, and F. Soldovieri, "Archaeology and cultural heritage," in *Subsurface Sensing*, A. S. Turk, K. A. Hocaoglu, and A. A. Vertiy, Eds. Hoboken, NJ, USA: Wiley, 2011, pp. 644–667.

[6] J. R. Butnor *et al.*, "Utility of ground-penetrating radar as a root biomass survey tool in forest systems," *Soil Sci. Soc. Am. J.*, vol. 67, no. 5, pp. 1607–1615, 2003.

[7] J. A. Doolittle and J. R. Butnor, "Soils, peatlands, and biomonitoring," in *Ground Penetrating Radar Theory and Applications*, 1st ed., H. M. Jol, Ed. Amsterdam, The Netherlands: Elsevier, 2009, pp. 179–202.

[8] J. J. Degenhardt and J. R. Giardino, "Subsurface investigation of a rock glacier using ground-penetrating radar: Implications for locating stored water on Mars," *J. Geophys. Res. Planet.*, vol. 108, no. E4, p. 8036, 2003.

[9] J. Francke, "Applications of GPR in mineral resource evaluations," in *13th Int. Conf. Ground Penetrating Radar*, 2010, pp. 1–5.

[10] A. Yarovoy, "Landmine and unexploded ordnance detection and classification with ground penetrating radar," in *Ground Penetrating Radar Theory and Applications*, H. M. Jol, Ed. Amsterdam, The Netherlands: Elsevier, 2009, pp. 445–478.

[11] D. J. Daniels, "Ground penetrating radar for buried landmine and IED detection," in *Unexploded Ordnance Detection and Mitigation*, J. Byrnes, Ed. New York, NY, USA: Springer, 2009, pp. 89–111.

[12] T. G. Savelyev, L. van Kempen, H. Sahli, J. Sachs, and M. Sato, "Investigation of time–frequency features for GPR landmine discrimination," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 118–129, 2007.

[13] A. Saintenoy and J. W. Hopmans, "Ground penetrating radar: Water table detection sensitivity to soil water retention properties," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 4, pp. 748–753, 2011.

[14] W. Shao, A. Bouzerdoum, S. L. Phung, L. Su, B. Indraratna, and C. Rujikiatkamjorn, "Automatic classification of ground-penetrating-radar signals for railway-ballast assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3961–3972, 2011.

[15] L. Liu and S. Liu, "Remote detection of human vital sign with stepped-frequency continuous wave radar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 775–782, 2014.

[16] A. M. Zoubir, I. J. Chant, C. L. Brown, B. Barkat, and C. Abeynayake, "Signal processing techniques for landmine detection using impulse ground penetrating radar," *IEEE Sens. J.*, vol. 2, no. 1, pp. 41–51, 2002.

[17] J.-H. Kim, S.-J. Cho, and M.-J. Yi, "Removal of ringing noise in GPR data by signal processing," *Geosci. J.*, vol. 11, no. 1, pp. 75–81, 2007.

[18] T. M. Grzegorczyk, B. Zhang, and M. T. Cornick, "Optimized SVD approach for the detection of weak subsurface targets from ground-penetrating radar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1635–1642, 2013.

[19] W. Shao, A. Bouzerdoum, and S. L. Phung, "Sparse representation of GPR traces with application to signal classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3922–3930, 2013.

[20] A. C. Gurbuz, J. H. McClellan, and W. R. Scott, "A compressive sensing data acquisition and imaging method for stepped frequency GPRs," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2640–2650, 2009.

[21] A. S. Venkatachalam, X. Xu, D. Huston, and T. Xia, "Development of a new high speed dual-channel impulse ground penetrating radar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 753–760, 2014.

[22] V. C. Chen and H. Ling, *Time–Frequency Transforms for Radar Imaging and Signal Analysis*. Norwood, MA, USA: Artech House, 2002.

[23] A. Papandreou-Suppappola, Ed., *Applications in Time–Frequency Signal Processing*. Boca Raton, FL, USA: CRC Press, 2003.

[24] Y. Sun and J. Li, "Adaptive learning approach to landmine detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 3, pp. 973–985, 2005.

[25] I. L. Al-Qadi, W. Xie, and R. Roberts, "Time–frequency approach for ground penetrating radar data analysis to assess railroad ballast condition," *Res. Non-Destr Eval.*, vol. 19, no. 4, pp. 219–237, 2008.

[26] W.-L. Lai and C.-S. Poon, "GPR data analysis in time–frequency domain," in *14th Int. Conf. Ground Penetrating Radar*, Shanghai, 2012, pp. 362–366.

[27] S. Sinha, P. S. Routh, P. D. Anno, and J. P. Castagna, "Spectral decomposition of seismic data with continuous-wavelet transform," *Geophysics*, vol. 70, no. 6, pp. 19–25, 2005.

[28] J. L. Starck, F. Murtagh, and J. M. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[29] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.

[30] O. Scherzer, Ed., *Handbook of Mathematical Methods in Imaging*, 1st ed. New York, NY, USA: Springer, 2011.

[31] J. Bobin, J. L. Starck, and R. Ottensamer, "Compressed sensing in astronomy," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 5, pp. 718–726, 2008.

[32] Z. Jindong, Z. Daiyin, and Z. Gong, "Adaptive compressed sensing radar oriented toward cognitive detection in dynamic sparse target scene," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1718–1729, 2012.

[33] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1200–1210, 2012.

[34] F. Fazel, M. Fazel, and M. Stojanovic, "Random access compressed sensing for energy-efficient underwater sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1660–1670, 2011.

[35] S. Pudlewski, A. Prasanna, and T. Melodia, "Compressed-sensing-enabled video streaming for wireless multimedia sensor networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 6, pp. 1060–1072, 2012.

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[37] J. Ma and F. X. Le Dimet, "Deblurring from highly incomplete measurements for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 792–802, 2009.

[38] V. H. Tang, A. Bouzerdoum, and S. L. Phung, "Two-stage through-the-wall radar image formation using compressive sensing," *J. Electron. Imaging*, vol. 22, no. 2, pp. 021006-1–021006-10, 2013.

[39] A. B. Suksmono, E. Bharata, A. A. Lestari, A. G. Yarovoy, and L. P. Ligthart, "Compressive stepped-frequency continuous-wave ground-penetrating radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 665–669, 2010.

[40] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, 2011.

[41] L. Stankovic, "A method for time–frequency analysis," *IEEE Trans. Signal Process.*, vol. 42, no. 1, pp. 225–229, 1994.

[42] I. Orović, S. Stanković, and M. Amin, "A new approach for classification of human gait based on time–frequency feature representations," *Signal Process.*, vol. 91, no. 6, pp. 1448–1456, 2011.

[43] L. Stankovic, T. Thayaparan, and M. Dakovic, "Signal decomposition by using the S-method with application to the analysis of HF radar signals in sea-clutter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4332–4342, 2006.

[44] D. Liu, Y. Zhao, B. Yang, and J. Sun, "A new motor fault detection method using multiple window S-method time–frequency analysis," in *Int. Conf. Syst. Informat.*, 2012, pp. 2563–2566.

[45] S. Mallat, *A Wavelet Tour of Signal Processing*. New York, NY, USA: Academic, 1999.

[46] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[47] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, 2007.

[48] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[49] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer, 2000.

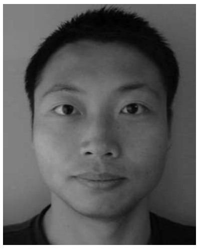
[50] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 1st ed. Cambridge, MA, USA: MIT Press, 2001.

[51] S. Abe, *Support Vector Machines for Pattern Classification*. New York, NY, USA: Springer, 2005.

[52] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[54] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Boston, MA: Morgan Kaufman, 2005.



**Wenbin Shao** received the B.Eng. degree in communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2003, and the M.Eng. and Ph.D. degrees in electrical engineering from the University of Wollongong, Wollongong, Australia, in 2007 and 2013, respectively.

His research interests include machine learning, pattern recognition, image and signal processing, and ground penetrating radar.

**Abdesselam Bouzerdoum** (M'89–SM'03) received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, WA, USA.

In 1991, he joined the University of Adelaide, Adelaide, Australia, and in 1998, he was appointed as a Associate Professor with Edith Cowan University, Perth, Australia. Since 2004, he has been a Professor of Computer Engineering with the University of Wollongong, Wollongong, Australia, where he also served as the Head of School of Electrical, Computer and Telecommunications Engineering (2004–2006) and Associate Dean of Research (2007–2013). He served as Deputy Chair of the Engineering, Mathematics, and Informatics Panel at Australian Research Council College of Experts from 2010 to 2011. He has published over 300 technical articles and graduated many Ph.D. and Research Masters students.

Dr. Bouzerdoum was a member of the Australian Research Council College of Experts from 2009 to 2011. He served as an Associate Editor of a number of international journals, including the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS (1999–2006). He was the recipient of numerous awards and prizes, including the Eureka Prize for Outstanding Science in Support of Defence or National Security in 2011, the Chester Sall Award in 2005, and a Distinguished Researcher Award (Chercheur de Haut Niveau) from the French Ministry of Research in 2001.

**Son Lam Phung** (M'02) received the B.Eng. and Ph.D. degrees from Edith Cowan University, Perth, Australia, all in computer engineering in 1999 and 2003, respectively.

He is currently a Senior Lecturer with the School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Wollongong, Australia. His research interests include the areas of image and signal processing, neural networks, pattern recognition, and machine learning.

Dr. Phung received the University and Faculty Medals in 2000.