Railcar Diagnostics Using Minimal-Redundancy Maximum-Relevance Feature Selection and Support Vector Machine Classification

Parham Shahidi[1], Daniel Maraini[1], and Brad Hopkins[1]

[1]*Amsted Rail Company, Inc. Chicago, IL, 60606, USA*

pshahidi@amstedrail.com
dmaraini@amstedrail.com
bhopkins@amstedrail.com

Abstract

Railcar condition is an important factor in the complex web of relationships between railroads, railcar leasing companies, shippers and railcar builders. The most important reasons for this are operational safety and economic considerations pertaining to equipment maintenance. In this study, an approach is presented for the diagnostics of railcar component health from vibration data, utilizing mutual information (MI) based minimal-redundancy-maximal-relevance (mRMR) feature selection and multi-class support vector machine classification. The proposed monitoring solution is a data-driven method which was developed with measurements taken at a railroad test laboratory under controlled conditions. Vibration data was collected from multiple locations on a railcar over several test runs, each utilizing wheelsets with different levels of wear. The input of controlled wheel wear levels was aimed at varying the system outputs to resemble those of cars with different levels of mileage in revenue service. The measured data sets were processed in the time domain, frequency domain and through wavelet transforms, resulting in the extraction of a set of 687 features from the acceleration signals. A maximum-relevance minimum-redundancy feature selection algorithm was used to find the optimal combination of features for classification. The algorithm performance was tested for the effect of feature set size, different kernels and scaling techniques on classification accuracy. The results and methods of this assessment are presented in the paper. The paper concludes with a proposal for a monitoring strategy aimed at specifically detecting faulty components and practicing predictive maintenance.

1. Introduction

The present work has the goal to develop a methodology for effective monitoring of freight rail bogies. It is motivated by a need in the freight rail industry to decrease asset maintenance related downtimes and to improve the effectiveness of maintenance schedules. The present study proposes using structured sensor data to monitor the health of the freight rail bogies through machine learning algorithms which pre-process the data, find the most relevant, non-redundant features and then make a classification decision. While the approach is a combination of existing techniques, it has not been applied to freight rail application before, making this a technique with the potential to modernize current railroad maintenance practices. This aspect is further emphasized by using domain expertise to select design parameters and ensuring real application constraints, such as power budget consciousness for on-board monitoring, were considered.

In (Shahidi, Maraini, Hopkins, & Seidel, 2014) the viability of applying on-board condition monitoring and diagnostics methods to freight rail applications was investigated and a framework to apply condition monitoring methods to freight rail bogies was established. The focus of the present study remains on the bogie as this is the component of a freight railcar which experiences the most wear and is most susceptible to fault modes. Figure 1 shows a standard North American three-piece bogie.

The trade association tasked with rule-making for freight rail transportation, the Association of American Railroads (AAR), has established a set of performance metrics (AAR, 2007) which all bogies have to meet before they can be deployed in service.

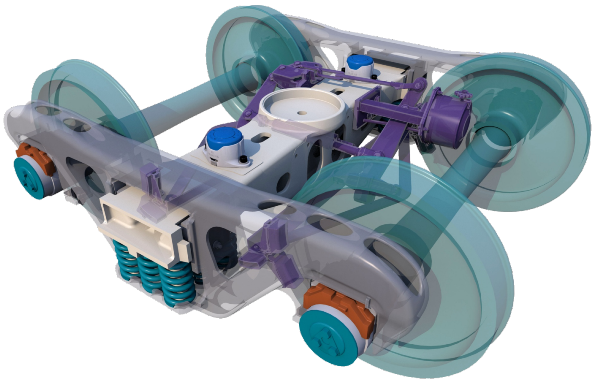


Figure 1. Standard North American three-piece bogie

After bogies go into service, maintenance is performed either as fixed-schedule or reactive maintenance. In the first case, maintenance downtimes are mostly avoided at the cost of unused capacity and premature component replacements. In the second case, wayside detectors, which are typically installed on the track, monitor passing railcars (Zakharov & Zharov, 2005) and generate need-based maintenance alerts. The two most common types of wayside detectors for railcar bogie performance are Truck Performance Detectors (TPD) and Truck Hunting Detectors (THD)[1]. Both of these detectors consist of strain gage based instrumentation which is added to the track to measure the lateral and vertical forces that railcar wheels exert on the rail. TPDs achieve this through instrumentation of two reverse curves with strain gages to measure the wheel lateral and vertical forces and wheelset angle of attack during curving. THDs use strain gages that are placed on tangent track to measure lateral wheelset oscillations. As of 2013, approximately 15 TPDs and 172 THDs were in service across the 140,000 miles of North American rail network. Other, even less common types of wayside equipment include Acoustics Bearing Detectors (ABD) and laser/vision-based systems. Deployment of these systems is in the low double digit numbers across the North American rail network and measurement results are often inconclusive (Tournay & Lang, 2007). The small number of detectors relative to the large size of the US rail network makes it clear that wayside detectors do not provide sufficient coverage and accuracy to comprehensively monitor freight rail bogie performance. The focus of wayside equipment on wheels indicates that the wheel rail interface is the most relevant research focus for investigations that pertain to railcar performance. Therefore the present study limits the evaluation of railcar bogie performance also to the effect of wheel wear on the performance.

2. RAILCAR DIAGNOSTICS

On-board condition monitoring is an area with large potential for research. As the name implies, the study combines the study of railroad engineering with the modern disciplines of diagnostics, prognostics and health management. Few studies with the same scope have been previously completed due to a number of reasons.

First, on-board condition monitoring has historically not been applied to freight rail applications and is a new technology in the realm of freight rail maintenance. Typically, condition monitoring in the freight rail industry is achieved through wayside equipment and therefore research in this area has traditionally focused on efficiency improvements. Barke and Chiu (Barke, 2005) published a review of existing freight rail bogie condition monitoring technologies but excluded on-board methodologies and solely focused on wayside technologies. Lagnebäck also limited his study of potential cost and efficiency improvements through condition monitoring (Lagnebäck, 2007) to wayside techniques, which resulted in recommendations to expand implementation.

Second, most condition monitoring studies have been attempted in the area of passenger rail transport (Ward, Goodall, Dixon, & Charles, 2010; Ward et al., 2011). Passenger rail bogies use complete and rigid frames and, therefore, do not have the issue of nonlinearities from the friction based suspension elements as three-piece bogies do. However, passenger bogies still have to deal with other nonlinearities, such as those from the wheel-rail interface. The difficulty of modeling nonlinearities was shown by Xia and True's study of nonlinear dry friction damping with hysteresis and stick-slip action in the friction forces on the contact surfaces of friction wedges (Xia & True, 2003).

Third, condition monitoring of freight rail applications is not limited to bogies and bogie suspension components. Other areas of interest, where significant work has been completed, include the wheel-rail interface (Hubbard, Ward, Goodall, & Dixon, 2013), railcar speed inaccuracies due to stick-slip action (Mei & Li, 2008), end-of-car devices (Hopkins, Seidel, Maraini, & Shahidi, 2015) and on-board weighing (Maraini, Shahidi, Hopkins, & Seidel, 2014) applications. It is understandable that the emergence of on-board monitoring technologies and continuous improvements in accuracy lead to a vast scope of interest which includes monitoring strategies for components which have traditionally not been able to be monitored effectively.

With the high cost of both preventive and reactive maintenance, condition-based maintenance can be considered the best solution to the problem at hand. Typically, applications follow one of two paths: either that of

[1] In the context of railroading and for this paper, the terms bogie and truck can be used interchangeably.

model-based condition monitoring or that of data driven condition monitoring.

For model-based condition monitoring, a physics-based model, derived from first principles, is used to determine required system parameters. The system parameters are then compared against data to determine if a deviation from a healthy system state is taking place. In (Li & Goodall, 2004) this approach was used in a two degrees-of-freedom, half-vehicle bogie model to determine such parameter deviations. For the data driven case, a signal from the system under test is used to infer what the current system health is. The signal must have a causal relationship to the system component subject to monitoring and thus be indicative of the system's performance. First, the signal is pre-processed and frequency and time domain based features are extracted. In many cases, the number of features can grow large and advanced techniques for selecting those features that are most descriptive are required. Feature selection algorithms include mutual information (Maraini & Nataraj, 2015) for estimating the similarity of two signals and minimal-redundancy-maximal-relevance (Kappaganthu & Nataraj, 2011; Peng, Fulmi, & Ding, 2005) for selecting an optimal feature subset.

The signal features constitute the inputs to machine learning algorithms which attempt to classify the health state of the system. If a target class is specified with the measurements, the problem is classified as a supervised learning problem and if no target class exists, the problem is classified as an unsupervised learning problem. Popular machine learning algorithms include techniques such as neural networks (Haykin, 1999) and support vector machines (Bishop, 2006; Cortes & Vapnik, 1995) to identify the fault modes from measurements.

In both cases, data is required to either compare against the model or to train the machine learning algorithm. Typically, this data is taken from inertial sensors such as accelerometers and gyroscopes, mounted on the system under test, but other metrics may be used as well. If prognostics is also part of the monitoring strategy, advanced filtering techniques such as particle filters (Arulampalam, Maskell, Gordon, & Clapp, 2002) or Kalman filters (Kalman, 1960) can be combined with the algorithm to estimate future states from the current state accelerometer measurements.

3. FIELD TEST

Data collection was conducted at Transportation Technologies Center, Inc. (TTCI) in Pueblo, CO. TTCI is a transportation research and testing organization which offers a wide range of tests for rail applications.

3.1. Field Test Setup

The Railroad Test Track (RTT) at TTCI, is a 13.5-mile loop with a primary purpose of high speed stability testing for excitation of lateral vehicle instability modes. The selection of lateral instability as the primary focus of this study was based on the fact that the main causes for this instability mode are the suspension parameters and wheel wear levels and thus directly influence the overall system performance.

For this study, one of the 50-minute (0.8 degree) curves was used to accelerate the train to target speeds ranging from 40 to 80 mph, broken up into approximately 5 mph increments per test run. Figure 2 shows the superelevation (upper plot) and curvature (lower plot) of the test track. Superelevation refers to the difference in height between the left and right rails (i.e., and indicator of cross-grade). Curvature refers to the degree to which the track deviates from completely straight.

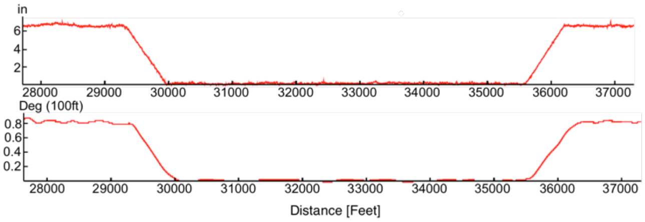


Figure 2. Test segment of RTT track

Once the target speed was reached, data acquisition systems began to measure accelerations at multiple locations on the car body and suspension. Test runs were aborted once either 80 mph or prescribed maximum acceleration limits were reached. The instrumentation setup included accelerometers with various dynamic ranges from ± 5 G to ± 200 G and gyroscopic sensors with rates of 250 °/sec. The sensor specifications were chosen to accommodate signal dynamic ranges that occur in the measurement locations. A HBM Somat eDAQ rugged data acquisition system was used to read the sensors with a sampling rate of 1000 Hz. The decision to use this sampling rate was based on knowledge of rigid and flexible modes of railcars experiencing lateral instability. Aliasing protection was ensured through analog filtering. Furthermore, it was observed that at elevated measurement locations on the carbody higher frequency content became attenuated. This is explicable through the behavior of the carbody, acting as a mechanical filter, which attenuates much of the frequency content above 10 Hz.

To test the system with controlled wear conditions as the inputs to the railcar system, wheels with three different levels of wear (new, medium and fully worn) were used. The individual wear levels were chosen as defined per AAR rules and the wheels were supplied at the test site by TTCI. For each round of testing the wheelsets were swapped out for sets with a higher degree of wear. Figure 3 shows the three wheel profiles, plotted against each other to visualize the effect of wear on the wheel profile geometry. The profile geometries shown in the figure are cross-sectional views of the wheel. The deviation in profiles occurring after 40 mm on the x-axis shows the portion of the profile in contact with the rail, which is subject to wear. The wear causes the wheels to develop a hollow profile on the running surface that is in contact with the rail, leading to unstable operating conditions. In (Klingel, 1883) the relationship between the level of wear and oscillation magnitude was provided. Every other aspect of the railcar and bogies remained unchanged to ensure that the wheel profiles were the sole factors influencing the stability of the railcar.

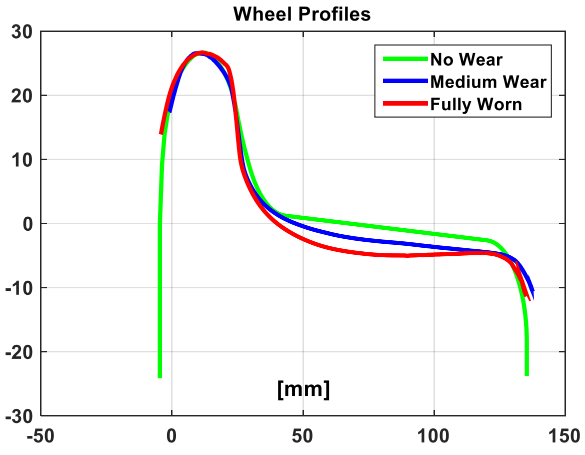


Figure 3. Different wheel wear profiles used as inputs

3.2. Field Test Procedure

As mentioned before, for each round of testing began an individual level of wheel wear was selected and run with a starting speed at or below 40 mph. The speed was then gradually increased in each run until the prescribed maximum acceleration limit per AAR regulations or a test speed of 80 mph was reached. The procedure was then repeated for the next level of wheel wear. Table 1 lists all combinations of wheel wear levels and test speeds that were evaluated. The green measurements indicate the speeds for the test runs which remained within the AAR limits for lateral instability and the red test speeds indicate where the limits were exceeded.

Table 1. Test speeds [mph] for each wheel wear level

No Wear	Medium Wear	Fully Worn
40	30	40
50	40	50
60	50	55
65	60	60
67	62	62
70	64	64
72		67
75		70

Figure 4 shows the vibration signals for the 64 mph runs for each wheel wear level. Overlaid in red is the averaged signal of each time series signal. It can be seen that the signals from figure 4 are reflective of the tabulated data. The vibration signal collected for the medium worn wheel (second subplot) has the highest vibration amplitude amongst the three signals. This corresponds with the test speeds from table 1, where the medium worn wheel set was run up to only 64 mph before lateral instability occured as shown between seconds 25 and 100, began.

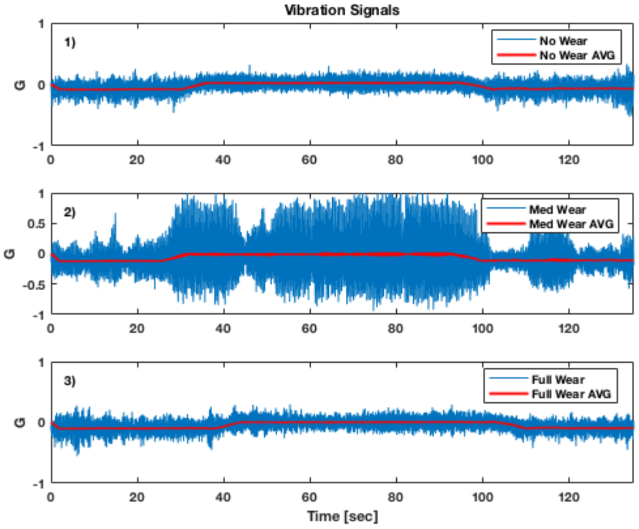


Figure 4. Vibration signals for three wear states at 64 mph.

The red signal in the plots above shows the effect of carbody tilt due to track super elevation on the zero position of the accelerometer which was calibrated on an even surface.

4. ANALYSIS

The analysis of the acceleration data was broken down into multiple subtasks which will be explained in this section. The first task was the extraction of the feature set from the data for each wheel wear state and test run. The feature extraction was further divided into extraction of time domain, frequency domain and wavelet transform features. Then, the data sets with the different wear states but same speeds were merged in a random sequence as the test signal. Since the wear levels for each sample time are known, this constitutes a supervised learning problem and a target class signal equal in length to the test signal was created. This was followed by partitioning the assembled data sets into training and validation sets. The training set was used to reduce the dimensionality of the feature matrix through a mutual information scheme which ranked the features and thereby allowed to exclude features with information gain below a user defined threshold. In order to minimize redundancy a minimum-redundancy-maximum-relevance algorithm was used next to refine the feature set. Then the reduced dimensionality training set was used to train a multiclass support vector machine. After training was complete, the validation set was used to evaluate the classification performance of the multiclass support vector machine in a one-versus-all classification scheme.

4.1. Feature Identification and Extraction

In the first analysis step, a set of features had to be identified for extraction and identification of faulty instability modes. In (Shahidi, Maraini, Hopkins, & Seidel, 2015) the feature set was identified as a combination of 14 features including the standard statistical moments, power content in three frequency bands, and two spectral measures. The three frequency bands were selected based on a qualitative spectrogram analysis in which the bands with the highest frequency content magnitude for faulty conditions were identified. In alignment with test results and the mathematical model of the oscillatory motion by (Klingel, 1883), the most important frequency band was chosen as the band between 2.5 and 3.5 Hz, which is the typical range for the track-damaging rigid body oscillation modes from the field test..

The choice of frequency bands for the analysis depends on the measurement location of the sensor and thus the component subject to performance monitoring. To accommodate different test setups, the above feature set was expanded to include a full frequency spectrum power band analysis between 0 and 500 Hz. The spectrum was divided into 5 Hz bands and each band was integrated individually to yield the band's power content. The individual power band spectral densities integration resulted in 100 additional features which cover the frequency bands for all components and can be combined with a feature selection algorithm to select the frequency bands with the highest discriminative power for classification for each location.

Table 2. List of Features

Feature #	Feature Description
1	Magnitude at Fund. Frequency
2	Fundamental Frequency
3	Mean
4	Variance
5	Standard Deviation
6	Peak to Peak
7	Skewness
8	Kurtosis
9	Hyperskewness
10	Hyperflatness
11	Crest Factor
12 . 111	5 Hz Power Bands between DC and 500 Hz
112 . 687	Features 3 – 11 computed on each level of a 64 level Wavelet Transform of the original signal

Since each test run typically lasted longer than 60 seconds and included non-stationary dynamic behavior of the carbody, a windowing approach was selected to compute the feature sets. A single five second data window incremented in one second intervals was selected as the best compromise between providing enough data for the feature computation, in particular the frequency resolution of the spectral analysis, and accommodating a reasonable monitoring framework. The latter is particularly important, considering that freight rail is typically not electrified and therefore requires power budget conscious on-board monitoring. The complete list of all 687 features is presented in table 2.

4.2. Cross Validation

After the features were extracted, the wheel wear states were assigned class labels $y \in \{1, 2, 3\}$, one for each level of wheel wear, and each sample/row of the feature matrix was associated with its corresponding class label. Then the labelled data sets, measured at the same speed, were assembled as a test sequence with random order. Figure 5 shows the target class. The order sequence of the class labels is medium wear, no wear, fully worn, medium wear.

A cross validation scheme was applied to the data to divide it into training and validation datasets. In prediction problems it is important to separate training and validation data to avoid overfitting and test generalization for independent datasets.

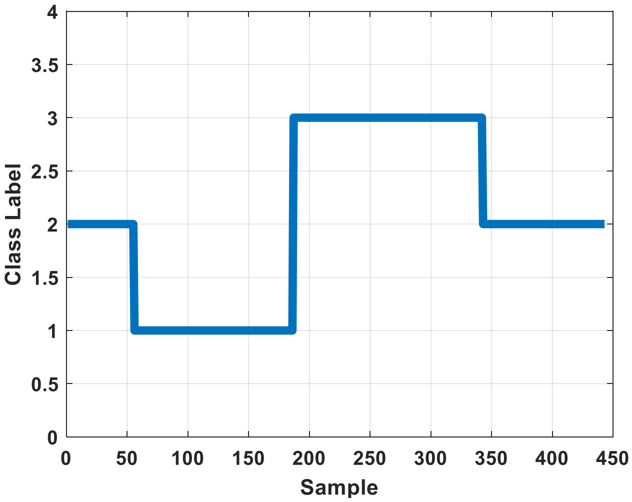


Figure 5. Three-class target class

The partitioning scheme was selected as a stratified hold out cross-validation which retained the proportions of the target class labels for the training and validation partitions. Additionally, to establish a repeatable accuracy, the scheme was reshuffled 10 times to provide additional validation data sets. The length of the validation partitions was selected as approximately one tenth the length of the original set.

4.3. Feature Selection Using Mutual Information

In cases with very large feature sets, a means to find and select only the most relevant features for the classification task is required to improve computational efficiency. Mutual information theory is a frequently used feature ranking algorithm to reduce the number of features. The idea is to compute a score which measures how informative each feature x_i is about the target class y. In other words, "*How much does a feature tell us about the target class?"*. The information provided by the algorithm can be used then to discard the features with the least amount of relevancy. Mutual information uses the entropy as the amount of information gain provided by each feature. Entropy is defined in eq. 1 as

$$H(X) = \sum_{x} p(x) \cdot \log \frac{1}{p(x)} \qquad (1)$$

where p_i is the probability of an event taking place with a certain outcome. An approximation of p_i for each feature can be obtained through the probability distribution of the scaled and discretized features. Discretization of the continuous accelerometer data is considered good practice to improve robustness of the various probability estimates which mutual information requires. The joint entropy of two random variables X and Y is defined in eq 2.

$$H(X,Y) = \sum_{x,y} p(x,y) \cdot \log \frac{1}{p(x,y)} . \qquad (2)$$

Together these quantities can be combined to calculate the mutual information (Peng et al., 2005) for each feature and the target class as shown in eq 3.

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \qquad (3)$$

Variations of eq. (3), based on axioms for different combinations of marginal, joint and conditional entropies, exist and can be used interchangeably.

A drawback of mutual information is that it selects only the most relevant features without taking the redundancy of the selected features into account. A computationally efficient extension of mutual information theory to address this shortcoming is *Minimal-Redundancy-Maximum-Relevance* (mRMR) feature selection as proposed by (Peng et al., 2005). In mRMR, the relevance of a feature subset is approximated through maximizing the mean of individual feature to target class mutual information.

$$\max D\,(S,c), D = \frac{1}{|S|^2} \sum_{x_i \in S} I(x_i;c) \qquad (4)$$

Conversely, minimal redundancy is achieved by excluding features with high individual feature to feature mutual information.

$$\min R\,(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i;x_j) \qquad (5)$$

The two measures in eqs. 4 and 5 are then combined through the $\Phi(D,R)$ operator and optimized to find the optimal feature subset in eq. 6.

$$\max \Phi(\mathrm{D}\,,\mathrm{R}), \Phi = \mathrm{D} - \mathrm{R} \qquad (6)$$

Table 3 shows the results of the optimization for a 10 feature subset of data collected at 65 mph.

Table 3. mRMR feature subset for 65 mph

Feature	Mutual Information
Band Power 205.0 – 209.9 Hz	0.9386
Band Power 90.0 – 94.9 Hz	0.7792
Band Power 210.0 – 214.9 Hz	0.8558
Fundamental Frequency	0.7799
Band Power 200.0 – 204.9	0.8338
Band Power 0.0 – 4.9 Hz	0.5300
Band Power 215.0 – 219.9	0.8668
Band Power 55.0 – 59.9	0.7372
Band Power 380.0 – 384.9	0.7821
Kurtosis	0.9274

In mRMR feature selection both the number of features and order of features are important. As implied by the algorithm, both are variables which change the resulting optimal subset.

It should be noted that since a stratified partitioning scheme was used in the algorithm, the results may slightly differ each time the algorithm is executed. The reason for this is that for stratification, samples are chosen from the population in no specific order as long as the overall proportion of the target class is maintained. Therefore single values can still vary under the same label and the variation this introduces may influence the probability distribution of the entropy calculation.

4.4. Multiclass Support Vector Machine Classification

A Support Vector Machine (SVM) is a maximum margin classifier that can be used for classifying both linearly separable and non-separable data. This is achieved by finding an optimal hyperplane which defines the maximum margin between two target classes. When the target classes are separable, the equation for the hyperplane is straightforward. However, for non-separable data, kernel based methods must be utilized to transform the data into a space whereby it becomes separable. In the case of classification with only two features a straight line can separate the target classes. However, when data with more than two features is to be separated, the line becomes a plane or hyperplane above 3 dimensions. The decision boundary is defined by eq. 7

$$y(\boldsymbol{u}) = \boldsymbol{w^T u} + b \qquad (7)$$

where ***y(u)*** is the decision, ***w*** a weight vector orthogonal to the decision surface, *b* a bias and ***u*** an unknown input vector. The optimal hyperplane can be found by solving the constrained optimization problem of the form given by

$$min \frac{1}{2} \|w\|^2 \qquad (8)$$

Limited by the constraint

$$t_i(\boldsymbol{w^T x_i} + b) \geq 1 \qquad (9)$$

In eq. (9), $\boldsymbol{x_i}$ represents known positive or negative training samples and $t_i \in \{-1,1\}$ is a factor that is either positive or negative depending on the sign of $\boldsymbol{x_i}$ so that (9) is always true. For the constraints, Lagrangian multipliers α_i are used to find the extremum of eq. (8). The Langragian which combines eq. (8) with the constraints from eq. (9) can be expressed as

$$L = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \left[t_i(\boldsymbol{w^T x_i} + b) - 1\right] \qquad (10)$$

Taking the derivative and setting it to zero gives the conditions for the extremum. Those can be plugged back into the original decision rule for a two-class classification problem of the form

$$y(\boldsymbol{u}) = \sum_i \alpha_i \, t_i \boldsymbol{x_i^T u} + b \qquad (11)$$

The vectors in the dot product in equation (11) can be transformed for cases when the classes are not linearly separable and in turn make them separable again. This is achieved using a kernel function of the form

$$\phi(\boldsymbol{x^T})\phi(\boldsymbol{u}) = k(\boldsymbol{x}, \boldsymbol{u}) \qquad (12)$$

For the present study all tests were conducted with a Radial Basis Function (RBF) kernel which is defined as

$$k(\boldsymbol{x}, \boldsymbol{u}) = exp\left(-\frac{\|x - u\|^2}{c}\right) \qquad (13)$$

and is applied to the unknown input vector ***u*** and known training examples ***x*** (Schölkopf et al., 1997).

The support vector machine is fundamentally a two-class classifier. To deal with the three-class separation problem of the present study with $y \in \{1,2,3\}$ in which each class label corresponds to one class for each wheel wear state, a multiclass support vector machine was used. A common approach for this is called the *one-versus-all* approach which constructs *K* separate SVMs in which the k^{th} model $Y_k(x)$ is trained using the data from class y_k as the positive examples and the data from the remaining *K-1* classes as the negative examples. For the present study, the result of this is a three support vector machine classification algorithm which is able to classify each individual wheel wear state against the other remaining wheel wear levels as a whole.

5. ANALYSIS RESULTS

The analysis was completed with the previously outlined algorithm and data from the field test. The primary focus of this study was to identify the three wheel wear states while testing for robustness of the algorithm against railcar speed. The secondary focus was to evaluate the influence of feature selection on classification accuracy. More precisely, to understand how many features are required for acceptable classification performance, and how beneficial mRMR is versus simple mutual information thresholding. The tertiary goal of the test was to evaluate the effects of various feature scaling techniques and improvements through more advanced kernel functions. For evaluation purposes, the main performance metric was classification accuracy, defined as the sum of true positives and true negatives divided by the total sum of samples.

Figure 6 shows the progression of the 10 mRMR selected features vs speed for each wheel wear level. The colors in figure 6 correspond with the colors of the wheel profiles in figure 3. The values for each feature are the average over a specific run (speed), therefore the horizontal axis label indicates the speed for that feature value. As presented in table 1, due to the experimental nature of the field data, the data sets for each fault were not always recorded at the exact same speeds. Hence, the features are also only available at the same speeds (as in table 1).

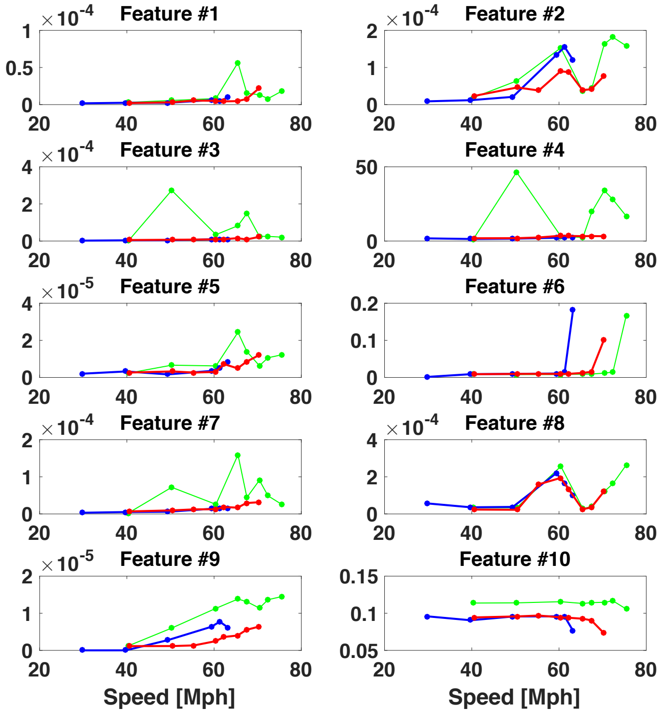
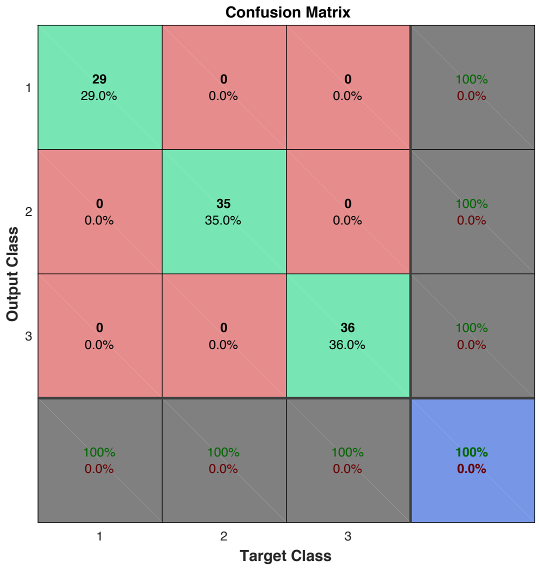


Figure 6. Progression of features versus speed – green stands for the no wear, blue for the medium wear and red for the full wear wheel profile.

Observation of figure 6 reveals an important attribute of the mRMR feature set: at higher speeds, approximately above 50 mph, the features tend to exhibit a discernible class separation. Especially for the test case at 65 mph, feature values of the new wheel (in green) show a clear separation from the medium (blue) and fully (red) worn wheel profiles. Since the primary focus of this study is the evaluation of the classification accuracy versus test speeds, the assembled data sets, corresponding to the target classes shown in figure 5, have to consist of data collected at the same speed for each level of wheel wear. Due to test constraints, only three test speeds (50, 60 and 65 Mph) were available for all wheel wear levels and are therefore also the only test speeds suitable for the analysis of the classification accuracy. For the first case, data from the 50 mph test run for each wheel wear state was used to evaluate the classification accuracy. The sequence of the wheel wear levels remained the same as presented in figure 5 and the feature set was selected as the optimized 10 feature mRMR set. After the first simulation with a hold-out cross validation scheme, 10 more simulations with a 10-fold cross validation scheme were run to find the average classification performance. The first (hold-out cross validation) simulation for the 50 mph test run yielded a classification accuracy of 99 % and the following 10-fold cross validation simulations also had an average accuracy of 99%. At 60 mph, the validation results were very similar with 100% for the hold-out cross validation and 99% for the 10-fold cross validation. Lastly, the 65 mph run again replicated the results from the 50 and 60 mph run with 100% and 99% accuracy for the hold-out and 10-fold cross validations respectively. Since the results for all speeds were similar, figure 7. only shows the confusion matrix for the 65 mph run with the optimized 10 mRMR feature set.



Confusion Matrix

Output Class	1	2	3	
1	29 29.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	35 35.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	36 36.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Target Class

Figure 7. Confusion Matrix for 65 mph with top ten mRMR features

The confusion matrix shows a perfect result for the hold out cross validation without any misclassifications when the top ten mRMR ranked features were used. This was also confirmed by the ROC curve shown in figure 8

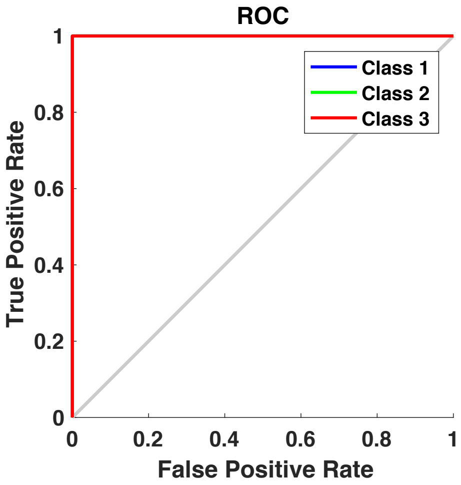
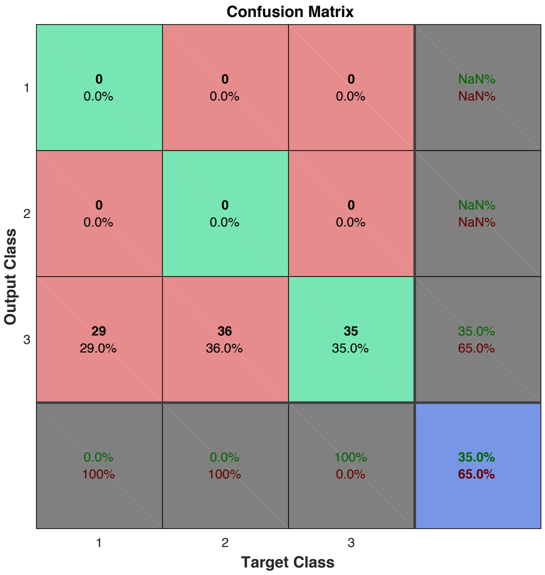


Figure 8. ROC curve for hold out cross validation at 65 mph with top ten mRMR features'

The secondary goal of this study was to evaluate the influence the number of features that were used and the feature selection algorithm has. For this purpose, the algorithm was reset first to exclude feature selection and thereby create a baseline scenario. This scenario yielded a hold-out validation accuracy of 35%, which was also reflected in the 10 fold cross validation. Interestingly, with this configuration, most samples were classified as fully worn as shown in the confusion matrix in figure 9.



Confusion Matrix

Output Class	1	2	3	
1	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
3	29 29.0%	36 36.0%	35 35.0%	35.0% 65.0%
	0.0% 100%	0.0% 100%	100% 0.0%	35.0% 65.0%

Target Class

Figure 9 Confusion matrix for 50 mph without feature selection

The achieved accuracy of 35% can thus be attributed to the distribution of class labels in the data set and means that the classifier is not able to separate the individual classes when too many features are used. The ROC curve shown in figure 10 confirms this as the lines are close to the diagonal, which means that pure chance without classification would have yielded the same results.

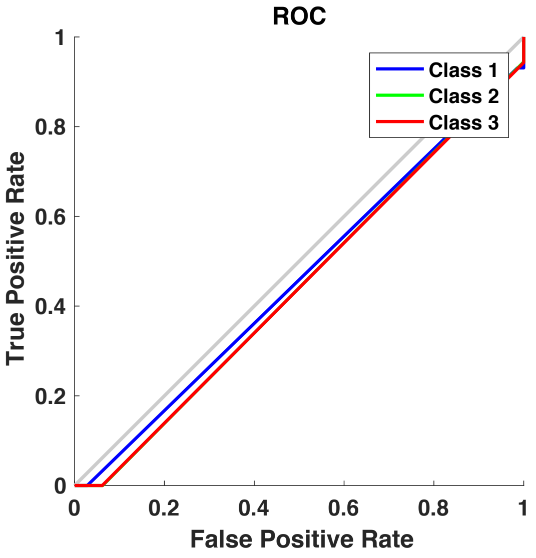


Figure 10. ROC curve for 50 mph without feature selection

To gain a deeper understanding of the classification accuracy as a function of the number of features that were used, a wrapper approach was used to iterate from the baseline to an optimal number of features. When configured as a wrapper, algorithms typically use the classifier as a "black box" to quantify the performance of the feature selection algorithm by using the classification accuracy as the performance metric. For this approach, each run of the classification module requires an individual set of data with training and test data subsets. To meet the secondary goal of the study, the input sets for the wrapper were selected as variations of the number of input features for each feature selection method. For the mutual information feature selection method, a 200 element array over the entire range of entropy values (0.2 to 0.938) of Mutual Information algorithm output was selected as the thresholds for the wrapper. For the mRMR feature selection method, the number of features in the optimal subset was varied through a 200 element array of number ranging from 1 to 500 features. With this configuration, the algorithm was instructed to execute for each input setting while measuring also the computational cost. The description of the wrapper approach in pseudocode is as follows:

```
for i:number of features
        divide samples into test and training sets
        train svm with MI_features(1:i)
        test svm with MI_features(1:i)
        train svm with mRMR_features(1:i)
        test svm with mRMR_features(1:i)
        calculate accuracies as TP+TN/(Total Samples)
        save accuracies in arrays
end
plot the accuracies vs number of features
```

Figure 11 shows the results for comparison of mutual information versus mRMR feature selection in form of a plot of the accuracies as they were calculated by the wrapper.

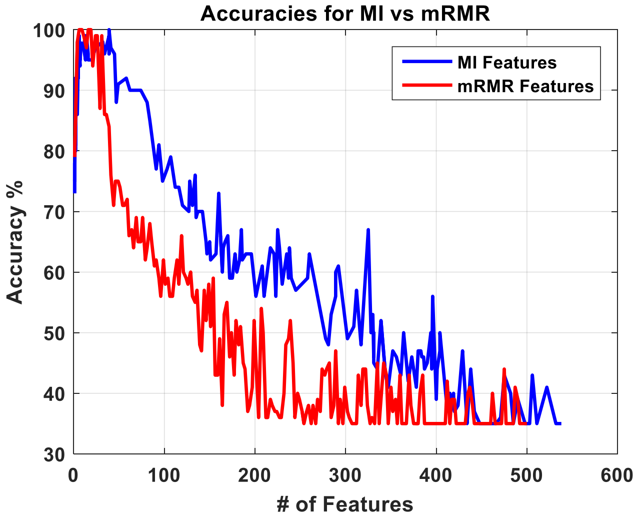


Figure 11. Comparison of MI and mRMR feature selection algorithms for different numbers of selected features.

It can be observed that for the majority of test cases mutual information feature selection typically yielded higher classification accuracy than mRMR feature selection, when equal numbers of features were selected. Furthermore, the baseline classification accuracy of 35% was also observed with very high feature numbers (above 200) which confirmed the earlier results from tests with complete set of 678 features. However, with very low numbers of features this trend changed. Figure 12 shows the zoomed section of figure 11 for low feature numbers. It can be seen that with up to approximately 20 features, mRMR feature selection yields higher classification accuracy to varying degrees. In the test case with only four features, the classification accuracy difference reached a value of 12%.

For the tertiary goal of this study, the wrapper was used to test four different feature normalization techniques. Normalization of the feature set is required to ensure that the support vector machine does not assign excess weight to one feature versus another. The four scaling methods tested in this study were no scaling, individual feature scaling (as compared to normalizing the feature set as a whole), scaling by the speed and combined (speed and individual) feature scaling.

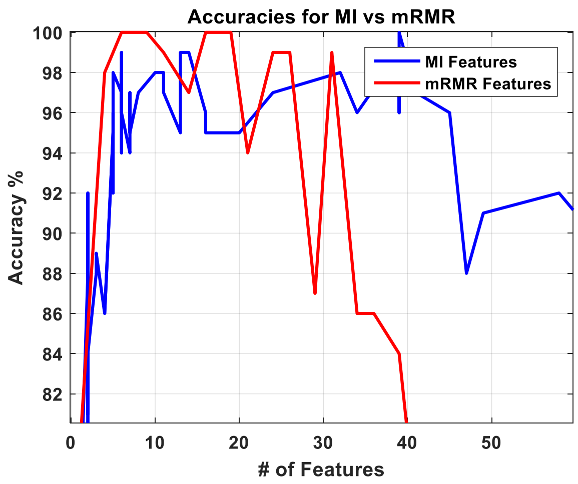


Figure 12. Zoomed in section of of MI and mRMR feature selection for different numbers of selected features.

The method for scaling was always subtraction of the mean and division by the standard deviation, except for scaling by speed. In individual feature scaling each feature was normalized by its own mean and standard deviation, whereas in scaling of the feature set as a whole, each feature was divided by the set's mean and standard deviation. Speed scaling was included to explore if the effect of small fluctuations (between 1-2 mph) of the train's test speed had affected the sensor measurements to a significant degree. Since speed was recorded as an instantaneous metric, each feature sample was divided by the instantaneous speed. It should be kept in mind that speed scaling was included as an exploratory scaling technique to see if a significant influence of speed flucttutations can be observed.

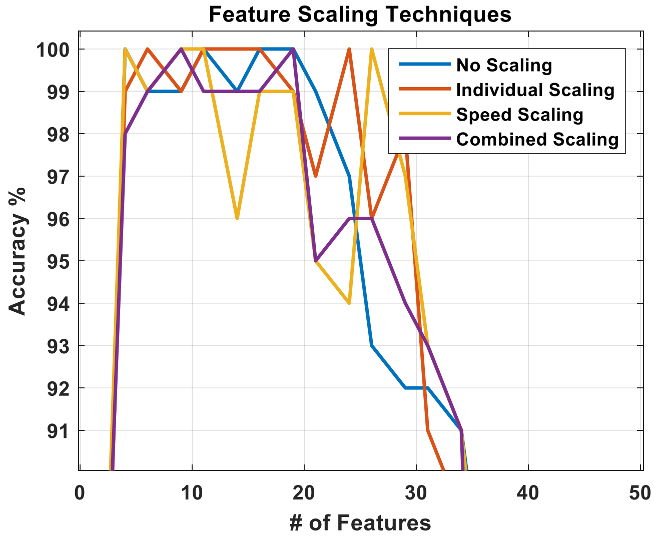


Figure 13. Evaluation of different feature scaling techniques

Lastly, a combined approach was also tested to understand the effect of speed and individual feature scaling simultaneously. For all four test cases, the mRMR algorithm was used as the feature selection algorithm based on the results from above. The results of this part of the analysis are shown in figure 13. It can be seen that the various scaling techniques only had little effect on the classification accuracy. The classification accuracies ranged between 99 and 100% in most cases with up to 20 features and thus it can be concluded that all four techniques yield approximately similar results.

As part of the tertiary goal of this study, a comparison of linear and nonlinear kernel techniques for the support vector machine was conducted. The result for this analysis are presented in figure 14. It can be observed that with a feature number that is less than 10, the RBF kernel has a better performance than the linear kernel. As the feature number increases, the trend reverses and the linear kernel exhibits better performance.

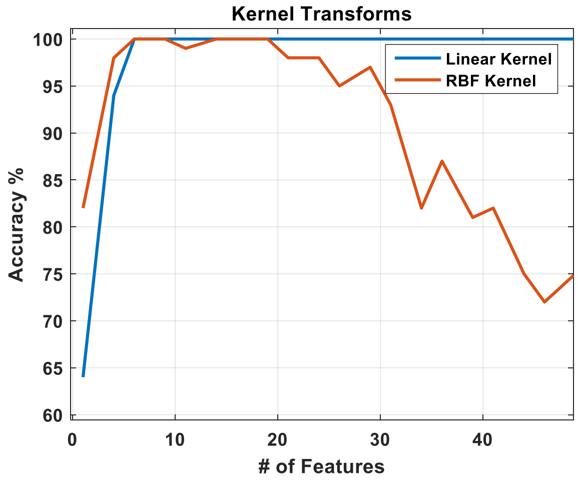


Figure 14. Comparison of linear and RBF kernel transforms for various numbers of features.

This result is telling as it shows one of the typically expected behaviors of support vector machines. It is well known in the literature (Hsu, Chang, & Lin, 2003), that with feature numbers n larger than the sample number m, so that $n \geq m$, a linear kernel typically performs better. This can be observed in this study as well, where even with a feature number of 10 the RBF kernel classification accuracy begins to deteriorate and the linear kernel accuracy remains stable. In the opposite case, where $n \leq m$, a transformation of the data into a higher dimensional space can provide better accuracy. This is also reflected in figure 14 where it can be seen that with less than 10 features the RBF kernel delivered better performance than the linear kernel. Another aspect is the computational efficiency which was approximately 50% higher for the linear kernel than for the RBF kernel. Thus, computational power, which is important for power budget conscious computation, as mentioned in section 4.1, as well as the number of selected features significantly influence which type of kernel should be chosen.

In the beginning of the analysis section, the robustness of the algorithm performance versus the three test speeds was discussed. To continue that discussion, the wrapper approach for the three test speeds was used again to test for the relationship between feature numbers and test speed. The results are shown in figure 15 where it can be observed that the algorithm's performance remained robust over the three test speeds and that the accuracy remained between 95 and 100% up until 25 features in the feature subset.

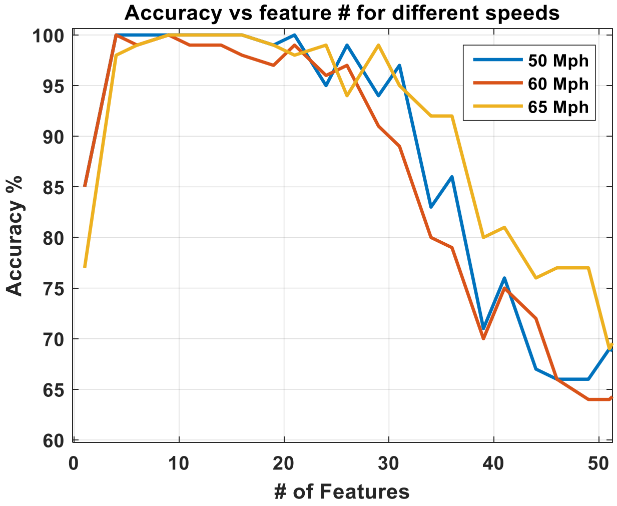


Figure 15 Comparison of classification accuracies for different speeds and feature numbers.

6. DISCUSSION

The analysis for the detection of wheel wear states from the vibration signature of acceleration data taken on the railcar was completed. In pursuit of the primary goal of this analysis, a success rate of 99% was achieved for the three test cases subject to testing over all test speeds. The high classification accuracy is mainly due to an extensive focus on generating adequate features and applying thorough feature selection techniques. Furthermore, the use of a RBF kernel in conjunction with a condensed and optimized feature subset was proven to deliver favorable results.

A few interesting points emerged from the analysis which require a deeper discussion. For the primary analysis goal, the large feature set size of 678 features has significantly broadened the breadth of pertinent information captured from the structured sensor data. This circumstance in conjunction with the feature selection algorithm contributed significantly to improved performance and high classification accuracies with robustness against speed fluctuations. Particularly the fact that the algorithm performance remained robust at lower speeds is noteworthy. Even though most of the features were sequentially (and thus automatically) generated, a deep frequency domain and wavelet transform analysis expanded the information from the original vibration signal significantly, allowing the full potential of mRMR to be utilized.

For the secondary goal, the investigation of *mutual information* versus *minimal-redundancy-maximal-relevance* feature selection, through a wrapper approach revealed that while MI will have better performance with high feature numbers, mRMR provides higher performance when few features are utilized. This is an important factor. In practical applications, keeping the number of features low is imperative to enable computational efficiency and algorithm simplicity. With mRMR, the classification accuracy remained as high as 95% even with as few as 2 features. This is a clear indicator of the benefits of using maximum relevance and minimum redundancy to find the difference between the two and thus optimize the feature subset.

Lastly, the tertiary goals of the study were met by understanding that feature scaling beyond normalization of the input features carries little benefit in the analysis. However, the choice of the kernel transform ended up being dependent on the number of features used for the analysis. Since this study was working with a higher number of features than number of samples, a linear kernel worked better with the full feature set. However, since a low feature number is more desirable for efficiency, the benefit of using a RBF kernel dominated the results with less features and became most evident when fewer than 10 features were used.

7. CONCLUSION

On-board condition-based maintenance for freight rail applications remains an underdeveloped field for the application of machine learning techniques. The industry has a growing need for advanced techniques which should be addressed in conjunction with domain expertise. Past efforts were mainly focused on passenger rail and wayside detection technologies. In the present study, a previously developed algorithm with above 90% accuracy was further improved to reach over 99% accuracy. The main conclusion from this is then "What is the next logical step?" and the answer is clearly the incorporation of prognostics techniques. It is expected that with a causal, multi-body dynamics based system such as a typical North-American three piece freight rail bogie, favorable results can be achieved if proper prognostics techniques are applied.

REFERENCES

AAR. (2007). Design, fabrication, and construction of freight cars *Manual of Standards and Recommended Practices C-II* (Vol. [M-1001]).

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Transactions on, 50*(2), 174-188.

Barke, D. a. C. K. W. (2005). Structural health monitoring in the railway industry: A review. *Structural Health Monitoring, 4*(1), 81 - 94.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4): springer New York.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Haykin, S.,. (1999). *Neural Networks - A Comprehensive Foundation*.

Hopkins, B., Seidel, A., Maraini, D., & Shahidi, P. (2015). *End-of-car Device Condition Monitoring with Onboard Sensors*. Paper presented at the *ASME Joint Rail Conference*, San Jose, CA.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification.

Hubbard, P., Ward, C., Goodall, R., & Dixon, R. (2013). Real time detection of low adhesion in the wheel/rail contact. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 0954409713503634.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, 82*(1), 35-45.

Kappaganthu, K., & Nataraj, C. (2011). Feature Selection for Fault Detection in Rolling Element Bearings Using Mutual Information. *Journal of Vibration and Acoustics, 133*(6), 061001-061001. doi:10.1115/1.4003400

Klingel, W. (1883). Über den Lauf der Eisenbahnwagen auf gerader Bahn. *Organ für die Fortschritte des Eisenbahnwesens, 20*, 113-123.

Lagnebäck, R. (2007). *Evaluation of wayside condition monitoring technologies for condition-based maintenance of railway vehicles*: Luleå University of Technology Luleå.

Li, P., & Goodall, R. (2004). *Model-based condition monitoring for railway vehicle systems.* Paper presented at the Proceedings of the UKACC international conference on control, Bath, UK.

Maraini, D., & Nataraj, C. (2015). Freight Car Roller Bearing Fault Detection Using Artificial Neural Networks and Support Vector Machines. In K. J. Sinha (Ed.), *Vibration Engineering and Technology of Machinery: Proceedings of VETOMAC X 2014, held at the University of Manchester, UK, September 9-11, 2014* (pp. 663-672). Cham: Springer International Publishing.

Maraini, D., Shahidi, P., Hopkins, B. M., & Seidel, A. (2014). *Development of a Bogie-Mounted Vehicle On-Board Weighing System.* Paper presented at the 2014 Joint Rail Conference.

Mei, T., & Li, H. (2008). Measurement of vehicle ground speed using bogie-based inertial sensors. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 222*(2), 107-116.

Peng, H., Fulmi, L., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(8), 1226-1238. doi:10.1109/TPAMI.2005.159

Schölkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *Signal Processing, IEEE Transactions on, 45*(11), 2758-2765.

Shahidi, P., Maraini, D., Hopkins, B., & Seidel, A. (2014). *Estimation of Bogie Performance Criteria Through On-Board Condition Monitoring*. Paper presented at the Annual Conference of the Prognostics and Health Management Society 2014, Fort Worth, TX.

Shahidi, P., Maraini, D., Hopkins, B., & Seidel, A. (2015). *Railcar Bogie Performance Monitoring using Mutual Information and Support Vector Machines*. Paper presented at the Prognostics and Health Management Society 2015, San Diego, CA.

Tournay, H. M., & Lang, R. (2007). History and Teardown Results of Five Loaded Coal Cars Identified as Poor Performers while Passing accorss a Truck Performance Detector *R-985*. Washington, DC: Association of American Railroads/ Transportation Technologies Center, Inc.

Ward, C. P., Goodall, R. M., Dixon, R., & Charles, G. (2010). Condition monitoring of rail vehicle bogies.

Ward, C. P., Weston, P., Stewart, E., Li, H., Goodall, R. M., Roberts, C., . . . Dixon, R. (2011). Condition monitoring opportunities using vehicle-based sensors. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 225*(2), 202-218.

Xia, F., & True, H. (2003, 22-24 April 2003). *On the dynamics of the three-piece-freight truck.* Paper presented at the Rail Conference, 2003. Proceedings of the 2003 IEEE/ASME Joint.

Zakharov, S. M., & Zharov, I. A. (2005). Criteria of bogie performance and wheel/rail wear prediction based on wayside measurements. *Wear, 258*(7–8), 1135-1141. doi:http://dx.doi.org/10.1016/j.wear.2004.03.025

BIOGRAPHIES

Parham Shahidi is a Project Engineer in the Bogie Systems Engineering group at Amsted Rail Company, Inc. He holds a BS degree in Mechanical & Process Engineering from TU Darmstadt in Germany and a PhD in Mechanical Engineering from Virginia Tech. Parham has been working in the area of condition monitoring since 2007, starting with the development of a speech based fatigue estimation system for train conductors in graduate school. In 2011 he joined Amsted Rail where his major projects in the area of health monitoring include the development of systems for vehicle instability detection, and bearing condition monitoring. He has published more than 10 technical articles related to his research. He is also an adjunct professor at Virginia Commonwealth University's Mechanical Engineering department and a board member of the German Engineers Association (VDI) in North America.

Daniel Maraini is the Manager of Condition Monitoring Engineering for Amsted Rail Company, Inc. He leads a team of Project Engineers focusing on condition monitoring solutions for freight railcars. Daniel has been a part of Amsted Rail since 2008. In various positions, he has worked on wireless sensing, remote asset monitoring, and condition monitoring for freight rail applications. Daniel is currently a PhD student in the department of Mechanical Engineering at Villanova University. His research is in the field of machinery diagnostics, with a focus on nonlinear model based techniques for rolling element bearings. He holds a MS in Mechanical Engineering from Villanova University and a BS in Physics from West Chester University of PA.

Brad Hopkins is a Project Engineer in the Bogie Systems Condition Monitoring group at Amsted Rail and a Lecturer in the Mechanical Engineering Department at Southern Illinois University at Edwardsville. He holds a BS, MS, and PhD in Mechanical Engineering from Virginia Tech. He has been working on track and railcar condition monitoring since 2010, with a focus on accelerometer-based monitoring and algorithm development. His current work includes broken rail detection, wheel defect monitoring, and end-of-car system monitoring. His additional research interests are vehicle dynamics, modeling, and simulation, controls, and vibration analysis.