The Information Theoretic Approach to Signal Anomaly Detection for Cognitive Radio

Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2010, Article ID 740594, 18 pages
doi:10.1155/2010/740594

Research Article

The Information Theoretic Approach to Signal Anomaly Detection for Cognitive Radio

Mostafa Afgani, Sinan Sinanović, and Harald Haas

The University of Edinburgh, AGB, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK

Correspondence should be addressed to Mostafa Afgani, m.afgani@ed.ac.uk

Received 1 December 2009; Accepted 13 March 2010

Academic Editor: Massimiliano Laddomada

Efficient utilisation and sharing of limited spectrum resources in an autonomous fashion is one of the primary goals of cognitive radio. However, decentralised spectrum sharing can lead to interference scenarios that must be detected and characterised to help achieve the other goal of cognitive radio—reliable service for the end user. Interference events can be treated as unusual and therefore anomaly detection algorithms can be applied for their detection. Two complementary algorithms based on information theoretic measures of statistical distribution divergence and information content are proposed. The first method is applicable to signals with periodic structures and is based on the analysis of Kullback-Leibler divergence. The second utilises information content analysis to detect unusual events. Results from software and hardware implementations show that the proposed algorithms are effective, simple, and capable of processing high-speed signals in real time. Additionally, neither of the algorithms require demodulation of the signal.

1. Introduction

Cognitive radio (CR) is the term used to describe smart, reconfigurable wireless communications devices that are capable of automatically adjusting their operating characteristics in order to adapt to changes in the radio environment. The purpose of such a system is to enable efficient use of the available radio spectrum and provide reliable service to the end user [2]. The motivation for efficient spectrum utilisation arises from the fact that it is a very limited resource. Although the electromagnetic spectrum is (for all intents and purposes) infinite, only a small fraction of it is usable for personal wireless communications as we know it today. Furthermore, while the spectrum available remains fixed, the number of wide-band wireless systems contending for access keeps growing—further compounding the spectrum scarcity problem.

Traditionally, the radio spectrum has been divided into a number of usable bands by regulatory bodies such as the Federal Communications Commission (FCC) in the USA and the Office of Communications (Ofcom) in the UK. Each of the bands is then assigned for exclusive access by a particular operator or service. A notable exception is of course the set of bands known as the industrial, scientific and medical (ISM) bands where emission from unlicensed consumer electronic devices is tolerated. While this restrictive approach to sharing the radio spectrum succeeds at providing a certain degree of interference protection, it is an inefficient use of the available resources since it is extremely unlikely that all of the bands are in use at the same time at a given place.

CR systems aim to simultaneously provide better quality of service and spectrum utilisation by dynamically moving the communication link from crowded or occupied bands to ones that do not appear to be in use by a primary licensed system at that instant. In order to carry out this task, secondary CR devices perform *spectrum sensing*—a procedure used to identify "holes" (free bands) in the spectrum and characterise the radio environment [3]. While there are a number of diverse approaches to problem [4], none of them are perfect. Energy detection-based methods [5] are limited by signal-to-noise ratio (SNR) constraints while methods relying on cyclostationary features [6] are limited by the amount of a priori information available regarding the signal structure of the primary system. As a result of these shortcomings, spectrum sensing cannot completely avert the risk of interference that arises from

dynamic spectrum sharing. Since interference generally leads to *anomalous* signal behaviour, an additional layer of simple signal processing algorithms that can help detect and characterise that behaviour is useful.

Anomaly detection refers to the process of locating unusual and unexpected events that may exist alongside nominal samples in a dataset. It is a process that is already utilised in a large number of diverse application domains. Typical examples include the detection of: unauthorised access to computer systems [7], irregularities in vital signs such as electrocardiogram (ECG) traces [8], fraud in financial services [9], and so forth. An extensive survey of current anomaly detection techniques and application domains is provided in [10].

The aforementioned survey reveals that there are many different approaches to solving the anomaly detection problem—each with its own set of advantages and disadvantages. However, there is one drawback that is shared by most algorithms: computational complexity. The computational effort required makes it difficult to adapt these techniques for real time and online processing of the input signal. This is unfortunate since any algorithm employed on an interactive communications system such as a CR platform must be capable of real time operation to maintain a seamless user experience. To overcome this challenge, two complementary anomaly detection algorithms based on simple information theoretic measures have been developed and are presented in this paper. The first method utilises Kullback-Leibler divergence (KLD) [11] while the latter uses the information content of individual signal events [12]. The algorithms are easy to generalise and broadly define anomalies as events that lead to changes in the nominal probability distribution of the radio signal. As a result, it is possible to employ the techniques for the detection of a wide range of disruptive events such as interference, timing errors, transmitter malfunction, and so on.

KLD is a convenient and robust method of measuring the difference between two data sets in a statistical sense. Due to its versatility and general appeal, it finds use in fields as diverse as economics [13] and computational neuroscience [14]. As a statistical comparison tool, KLD can also be employed for the automatic and real time detection of unusual (anomalous) data segments. The proposed KLD-based technique utilises two data windows to perform a statistical comparison of neighbouring segments of signals with periodic structures (e.g., systems utilising time division multiple access [15]). Since segments separated by the signal period are expected to be analogous and hence have similar statistical characteristics, any deviation can be taken to imply the presence of an anomaly.

Unlike the KLD-based method, the information content analysis (ICA) algorithm can also be applied to signals lacking any kind of periodic features. Information content is a quantity that is directly related to the probability of an event: the lower the probability, the higher the information content. Since anomalies are, by definition, rare (low probability), the associated information content is high. The proposed anomaly detection algorithm exploits this fact by analysing the signal for high-information content events.

Software implementations of the algorithms have been tested against a set of real wireless signals with promising results. Additionally, a Xilinx Virtex4 field-programmable gate array-(FPGA-) based hardware implementation of the KLD-based method has shown that the algorithm is indeed capable of real time analysis of high speed, high bandwidth signals.

A brief review of some of the anomaly detection algorithms described in literature is provided in Section 2 while the proposed algorithms are described in detail in Section 3. Results from applying the techniques to the test signals and measures of performance are provided in Section 4. The hardware implementation is briefly discussed in Section 5 while Section 6 concludes the paper with a summary of the contributions made and directions for future work.

2. Review of Existing Methods

Anomaly detection, also known as novelty detection or outlier detection, is a rich field of research with a very large body of work that exists in the literature. The existence of multiple survey-type papers such as [10, 16–20] is a testament to the true extent of the subject of anomaly detection. It is therefore surprising to learn that it is still very much an active area of research lacking generic algorithms that can be applied universally to anomaly detection problems. Most of the methods described in literature are based on tightly constrained frameworks that apply to very specific classes of problems.

Existing techniques of anomaly detection can be separated into a handful of classes depending on the underlying approach. *Classification*-based methods utilise supervised machine learning techniques to categorise nominal and anomalous behaviour while *clustering* and *nearest-neighbour* based techniques rely on measures of the relative distance between points of data. *Statistical* techniques detect anomalies by comparing the test data points against stochastic models of nominal behaviour. *Information theoretic* methods employ measures of information such as Kolmogorov complexity and entropy and work under the assumption that anomalies lead to a change in the information content. The algorithms proposed in this paper employ techniques that are both statistical and information theoretic in nature.

A statistical method of detecting anomalies in sensor data streams is proposed by Basu and Meckesheimer in [21]. Relying on the assumption that the data stream is continuous, the method exploits the fact that correlation between neighbouring data points is higher than between points separated by a relatively long length of time. The described algorithm detects anomalous events by comparing the value of each event against the median of a data set composed of neighbouring events. The performance of the method then depends on the size of the data set and the threshold. Since the algorithm expects an input where subsequent data points change little under nominal circumstances, it is unsuitable for use in typical communications systems where the signal strength can vary considerably even under normal operating conditions.

An algorithm for detecting anomalous network traffic by means of a combined statistical and information theoretic

measure is described by Krügel et al. in [7]. For each packet, an anomaly score is computed by considering the packet type, length, and payload distribution. If the combined score exceeds a certain threshold established through training, existence of an anomalous packet is signalled. Since the algorithm is designed for operation in the network layer, it cannot be utilised for link monitoring and anomaly detection in the physical layer.

Another set of statistical anomaly detection algorithms are presented by Desforges et al. and Yeung and Chow in [22, 23], respectively. Both papers propose the use of the Parzen windows method of nonparametric smooth probability density estimation in order to establish a stochastic model of the data distribution. While Yeung and Chow simply test whether a data point belongs to a given model, Desforges et al. also construct a model of the test data set and compare that against the reference. Since the model of the underlying process is determined once at the onset of the experiments, the algorithms cannot cope with nonstationary systems. Utilisation of Parzen windows method for density estimation also makes the algorithms computationally expensive and unsuitable for real time implementation.

A technique for detecting anomalous segments ("discords") in structured time series such as ECG traces is described by Lin et al. in [8]. Given a time series containing a discord, the algorithm essentially splits the series into a set of small segments and computes the mutual distance between the segments. If a segment is then found to have a minimum distance larger than a predefined threshold, it is labelled as anomalous. Although the algorithm shows promising results, it is unsuitable for real time implementation due to the computational complexity cost associated with performing a search for anomalous segments.

Finally, the use of various information theoretic measures for anomaly detection is discussed by Lee and Xiang in [24]. However, the focus of the paper is on determining the suitability of data models through the use of measures such as entropy and relative entropy (i.e., KLD) rather than algorithms for detecting anomalies.

It is evident from this survey of existing techniques that there is a lack of algorithms that offer the features needed (nonparametric with a low computational complexity and the ability to handle nonstationary behaviour) to analyse radio frequency signal envelopes in real time for anomalies.

3. Anomaly Detection Algorithms

The detection algorithms utilise KLD and information content analysis, respectively, to determine the presence of anomalies. Both quantities are ultimately calculated from estimates of the statistical probabilities of events in the signal.

Given two data sets P_n and Q_n, at time n, that contain samples from domain X, it is possible to obtain empirical estimates of the associated probability mass functions (PMFs) p_n and q_n from a nonparametric model such as a histogram. Once the PMF estimates are available, the KLD between them can be calculated using [11]

$$D(p_n \| q_n) = \sum_{x \in X} p_n(x) \log_2 \frac{p_n(x)}{q_n(x)}, \tag{1}$$

where $x \in X$. Since base-2 logarithm is used, the divergence is measured in *bits*. KLD between two PMFs is generally asymmetric: that is, $D(p_n \| q_n) \neq D(q_n \| p_n)$ and the triangle inequality is not satisfied. When $p_n = q_n$, the KLD is zero; otherwise, it is a positive real number ($\mathbb{R}_+$). For brevity and convenience, $D(p_n \| q_n)$ will also be referred to as D_n in this paper.

KLD belongs to a class of distance measures known as *f-divergence* (or *Ali-Silvey distances*). Some of the other distance measures that belong to the same class are variational distance (symmetric), Hellinger distance (symmetric), and Chernoff distance (generally asymmetric) [25]. While they are all equally suitable for quantifying the statistical difference between two probability distributions, KLD and variational distance are the least complex and therefore the easiest to implement. Variational distance is defined as

$$\begin{aligned} V(p_n \| q_n) &= \frac{1}{2} \sum_{x \in X} | \, p_n(x) - q_n(x) \, | \\ &= \frac{1}{2} || p_n - q_n ||_1, \end{aligned} \tag{2}$$

where $\| p_n - q_n \|_1$ is commonly known as the $\mathcal{L}_1$ distance (L1D) between the PMFs p_n and q_n. Furthermore, KLD and L1D (and hence the variational distance) are related by the inequality [11]

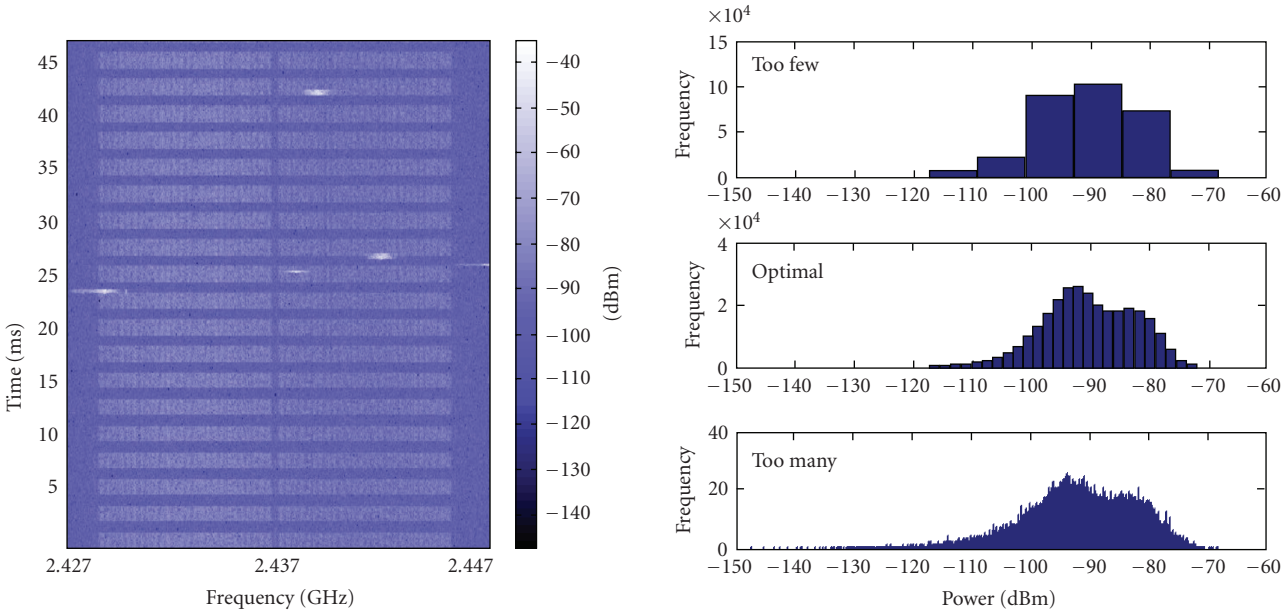
$$D(p_n \| q_n) \geq \frac{1}{2 \ln 2} || p_n - q_n ||_1^2. \tag{3}$$

Crucially, it states that $D(p_n \| q_n)$ is bounded by $\| p_n - q_n \|_1^2$ and not $\| p_n - q_n \|_1$. It is an important distinction as it implies that for certain PMF pairs the KLD may in fact be *smaller* than the L1D. For a pair of largely dissimilar PMFs (Differences that are large enough to produce a $\mathcal{L}_1$ distance of $2 \ln 2$ or greater, to be precise.), as is generally the case when comparing an anomalous data set against a nominal reference, larger distance magnitudes are obtained from KLD rather than L1D. However, when both PMFs are similar (e.g., a nominal data set and the reference), this can lead to L1D values that are larger compared to KLD—increasing the likelihood that false positives are detected. As a result, it is expected that KLD is better suited for statistical anomaly detection compared to L1D. This is confirmed by the results seen in Section 4 where the performance of a KLD-based algorithm is compared against one based on L1D. The algorithm for anomaly detection using KLD is described in Section 3.2.

Information content analysis is another technique based on an information theoretic quantity that can be utilised for the detection of anomalies. The amount of information, $I_n(x)$, conveyed by any discrete random event, x_n, at time n, is directly related to its probability of occurrence, $p_n(x)$ [12]:

$$I_n(x) = -\log_2 \{ p_n(x) \}, \quad n = 1, 2, \ldots \tag{4}$$

Since base-2 logarithm is used once again, information is also measured in *bits*. The equation implies that an event with a very high probability of occurrence carries very little information while a large amount of information is



(a) Signal with interference

(b) Effect of bin count

Figure 1: (a) Spectrogram of a wireless local area network (WLAN) signal experiencing interference from a Bluetooth device. The regular pattern is a single WLAN frame repeatedly transmitted by a signal generator. The frequency hopping nature of the Bluetooth transmission is clearly visible in the plot. (b) Impact of the number of bins utilised, β, on a histogram of the instantaneous power density of the WLAN signal. When β is too small, the histogram is insensitive to small changes and does not effectively capture the subtleties of the process. At the other extreme, it is too sensitive and therefore susceptible to noise. The optimal β is with respect to some minimum error criterion [1] and provides a good balance between resolution and sensitivity.

conveyed by the occurrence of rare events (i.e., $I_n(x) \to \infty$ as $p_n(x) \to 0$). Information is always real, positive ($\mathbb{R}_+$) and monotonically increasing with decreasing values of event probability. ICA is essentially a nonlinear scaling function that favours the unusual.

3.1. Histogram and PMF Estimation. It is clear from (1) and (4) that both KLD analysis and ICA require estimation of empirical event probabilities. One approach to obtaining the necessary estimates is via event histograms. In addition to being simple to implement, histograms are nonparametric—implying that no assumptions need to be made regarding the underlying distribution of the sample data.

For samples that originate from domain X, the histogram is obtained by first partitioning X into bins B such that

$$X = \bigcup_{l=1}^{\beta} B_l, \tag{5}$$

and then counting the number of samples that belong to each bin. β is the total number of bins used to construct the histogram. Once the histogram is available, the empirical PMF of the sample set is easily obtained by simply dividing the histogram by the cardinality of the set.

Given a statistically significant sample size, it is clear that the only parameter that affects the quality of the PMF estimate obtained is the bin allocation B. If the partitions are then assumed to be equidistant for simplicity, the only variable that remains is the number of bins utilised: β. The effect of β on the histogram of a random process is shown in Figure 1. The random process in question is the instantaneous power density at any time-frequency point of the signal shown in Figure 1(a). It is a wireless local area network (WLAN) signal experiencing bursts of interference from a Bluetooth (BT) device. Histograms of the power density obtained using three different values of β are shown in Figure 1(b).

When a small number of bins are utilised, that is, β is small, the histogram is insensitive to small scale variations in the input. As a result of the poor resolution, the estimated model fails to adequately capture the subtleties in the behaviour of the underlying random process. On the other hand, when the value of β utilised is too large, the resolution is too high and the histogram is overly sensitive—resulting in an estimate that is noisy. The optimal value of β yields a good balance between resolution and sensitivity.

A method of computing the optimal bin size (and hence the optimal β) for constructing a histogram, subject to some minimum mean square error criterion, is provided in [1]. While the algorithm described therein is conceptually simple, it unfortunately requires the use of exhaustive search to iteratively minimise a certain cost function—making it too computationally expensive to be evaluated in real time on a hand-held mobile device with limited energy and processing power.

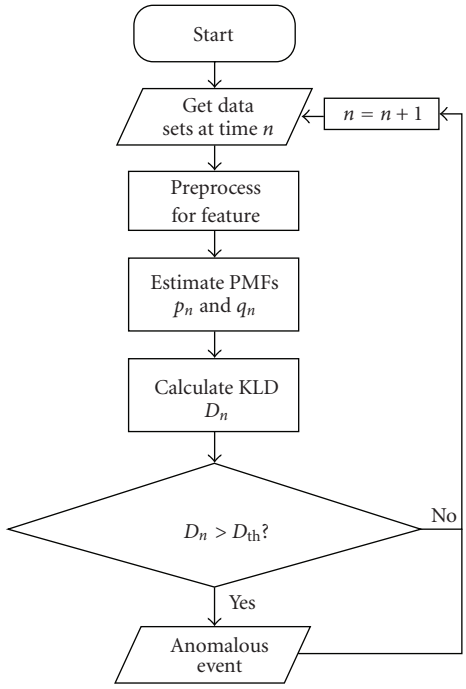


Figure 2: The algorithmic flowchart. KLD is used to compare the statistical distribution of a test data set against that of a reference. If the divergence, D_n, is greater than some predefined threshold, D_{th}, the test set may be anomalous.

The impact of β on the performance of each of the anomaly detection algorithms has been investigated and the results are presented in Section 4.2.4.

Depending on the choice of B, there may be zeros in the estimated PMFs due to the presence of empty histogram bins. Under such circumstances, calculation of the KLD using (1) can be a problem as it leads to instances where $0\log_2\{0/q_n(x)\}$ or $p_n(x)\log_2\{p_n(x)/0\}$ have to be evaluated. While it is certainly possible to handle these as special cases by setting them to 0 and, ∞, respectively, through continuity arguments, it may be better to simply avoid zeros in the PMFs. It is possible to avoid empty histogram bins and hence zeros in PMFs by adding a small number, λ, to every bin of the histogram. As preloading of histogram bins in this manner undoubtedly distorts the estimate of the true PMF, the preload value must be carefully chosen. According to the work done by Krichevsky and Trofimov [26] and Johnson et al. [27] $\lambda = 0.5$ is a good choice.

3.2. Algorithm Based on KLD. The capability of KLD to quantise the difference, in a statistical sense, between two data sets to single real value is ideal for use in anomaly detection since it provides a convenient detection metric. A general description of the algorithm is provided here while a discussion of the optimisations needed for an efficient hardware implementation is provided in Section 3.3 that follows.

The flowchart in Figure 2 shows the proposed algorithm. At time n, the process starts with the acquisition of the two data sets to be compared using KLD. One of the data sets is a reference (Q_n) while the other is the one under test (P_n). If the samples in the data sets do not directly represent the parameter of interest, they must be processed. Once the data sets have been suitably transformed, the associated PMFs $p_n(x)$ and $q_n(x)$ are estimated and used to compute the KLD, D_n. If D_n is then observed to be larger than some predefined KLD threshold, D_{th}, the test data set may be anomalous.

Table 1: Complexity analysis of KLD.

Operations	Σ	$\times$	$\div$	log	Total
	$2\lvert P_n\rvert + \beta$	β	3β	β	$2\lvert P_n\rvert + 6\beta$
Memory	$\lvert P_n\rvert + \psi + 2\beta$				

This general approach to detecting anomalies using KLD can be easily adapted for use with signals containing periodic structures. One example of such a signal is IEEE 802.16e wireless broadband (WiBro) which utilises time division duplexing (TDD) [28]. Periodic signals are expected to have statistics that are also periodic—implying that segments of the signal separated by the period, T_p, should have probability distributions that are very similar under normal circumstances. Therefore, by simply acquiring the data sets P_n and Q_n from two sliding signal windows of length T_w with centres separated by T_p, the proposed algorithm can be utilised for the detection of anomalies in periodic signals. KLD analysis can be performed on the signal envelope itself and as a result, demodulation is unnecessary and the only a priori information required by the algorithm is the signal period T_p.

While the steps required to compute the KLD are all simple and straightforward, the storage (data buffers) and the number of arithmetic operations required grow linearly with the size of input data sets. As these data sets can be very large when analysing high speed signals, it can easily lead to scenarios where it may not be possible to provide for the resources required by the algorithm. Analysis of the algorithm's complexity and memory requirements follows and is summarised in Table 1.

The input data sets P_n and Q_n themselves require a buffer capable of holding at least $|P_n| + \psi$ elements, where ψ is the number of samples corresponding to the signal period and $|P_n|(= |Q_n|)$ is the size of the data windows in samples. Only a single buffer is required for the input data since one of the data sets is essentially just a ψ-delayed version of the other in this case. Computing the frequency count over the bins (B), for the purpose of estimating the histograms, requires up to $|P_n|$ additions for each of the two windows. Once the histograms are available, the PMFs are obtained by dividing the frequency count in each of the β bins by $|P_n|$. Two buffers of size β each are then required to store the resulting PMFs. Computation of the KLD from the PMFs then requires a further β divisions, logarithms, multiplication, and addition, respectively.

3.3. Hardware Implementation. The analysis performed in the previous Section 3.2 reveals that a direct interpretation of the algorithm to hardware would be inefficient, inflexible,

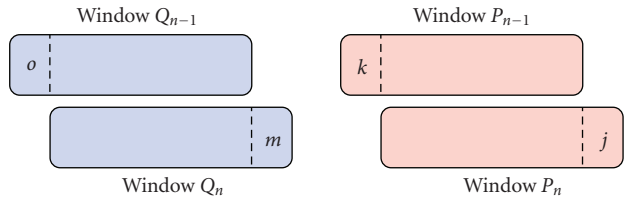


Figure 3: At any given time, only a maximum of four histogram bins need to be updated (two per window). Bins incremented (fresh samples) are denoted by j and m while bins decremented (old samples at the end of the window) are denoted by k and o.

Table 2: Complexity analysis of DKLD.

Operations	Σ	×	÷	log	Total
	20	8	16	16	60
Memory	$\lvert P_n\rvert + \psi + 2\beta$				

and computationally expensive. The inefficiency arises from the fact that at each successive time instance, the PMFs and the KLD are completely recalculated, even though it is only a single sample that changes in each of the data sets. The inflexibility comes from the fact that the computational complexity depends on β, implying that a direct interpretation would be limited by the initial choice of the histogram bin resolution. Finally, logarithms and divisions can be very costly to implement in hardware. Fortunately, there are several well-known methods that can be adopted to overcome each of these challenges.

The complexity that arises from the division operation in (1) can be removed by exploiting the identity $\log(a/b) = \log(a) - \log(b)$:

$$\begin{aligned} D_n &= \sum_{x\in X} p_n(x)\log_2\frac{p_n(x)}{q_n(x)} \\ &= \sum_{x\in X} p_n(x)\Big\{\log_2[p_n(x)] - \log_2[q_n(x)]\Big\}. \end{aligned} \tag{6}$$

The division operation is exchanged for a subtraction and a second base-2 logarithm operation which can be implemented in a very efficient manner by means of a lookup table.

Further efficiency improvements can be achieved by making application-specific changes to the way the algorithm is evaluated. Since the purpose of the algorithm is to analyse periodic signals by means of two sliding windows, it holds that at any given instance, only one sample in each of the data sets changes. This in turn implies that only a maximum of 4 histogram/PMF bins need to be updated at that instant—two for each data set/window. The two bins per window account for the freshly acquired sample (bin frequency count incremented by one) and the sample that is dropped at the end of the window (bin frequency count reduced by one). An illustration is provided in Figure 3.

This also means that the KLD values change very little between subsequent time steps for this particular application—suggesting that it is possible to rewrite (6) in the form of a differential equation:

$$\begin{aligned} D_n &= \sum_{x\in X} p_n(x)\Big\{\log_2[p_n(x)] - \log_2[q_n(x)]\Big\} \\ &= D_{n-1} \\ &\quad - p_{n-1}(j)\Big\{\log_2[p_{n-1}(j)] - \log_2[q_{n-1}(j)]\Big\} \\ &\quad + p_n(j)\Big\{\log_2[p_n(j)] - \log_2[q_n(j)]\Big\} \\ &\quad - p_{n-1}(k)\Big\{\log_2[p_{n-1}(k)] - \log_2[q_{n-1}(k)]\Big\} \\ &\quad + p_n(k)\Big\{\log_2[p_n(k)] - \log_2[q_n(k)]\Big\} \\ &\quad - p_{n-1}(m)\Big\{\log_2[p_{n-1}(m)] - \log_2[q_{n-1}(m)]\Big\} \\ &\quad + p_n(m)\Big\{\log_2[p_n(m)] - \log_2[q_n(m)]\Big\} \\ &\quad - p_{n-1}(o)\Big\{\log_2[p_{n-1}(o)] - \log_2[q_{n-1}(o)]\Big\} \\ &\quad + p_n(o)\Big\{\log_2[p_n(o)] - \log_2[q_n(o)]\Big\}, \end{aligned} \tag{7}$$

where the four bin indices j, k, m, and o are assumed to be unique. If not, any duplicate terms in the equation are set to zero.

The differential equation form of KLD (DKLD) shows that its computational complexity is no longer dependent on the number of histogram bins utilised in evaluating the PMFs. Assuming that D_{n-1} is available, only 16 additions/subtractions, 16 logarithms, and 8 multiplications are needed to calculate D_n—regardless of the value of β. This opens the path for a fixed complexity, flexible, and efficient implementation that can be easily updated to accommodate a wide range of histogram resolutions.

The computational complexity and storage requirement of DKLD are shown in Table 2. Comparisons against the unmodified, direct interpretation version of KLD (Table 1) reveals that while memory utilisation remains unchanged, there is a vast difference in the number of operations required. Regardless of the window size and histogram bin count, 60 operations are needed to compute the KLD. In addition to the 16 additions/subtractions, 16 logarithms, and 8 multiplications required for the DKLD (7), 4 more additions/subtractions are required to update the affected histogram bins and 16 divisions are required to obtain the necessary PMFs at times $n-1$ and n from the histogram.

Switching to a fixed-point representation and using a lookup table for base-2 logarithms provide further reductions in complexity at the expense of a slight increase in the memory requirements. The size of the table, L, then dictates the precision available. Additionally, if $\lvert P_n\rvert$ is chosen such that it is always a power of two (PoT), that is,

$$\lvert P_n\rvert = 2^{\sigma}, \quad \sigma = 0, 1, 2, \ldots, \tag{8}$$

no division operations are required to obtain the PMFs since division by a PoT is simply a bit-shift operation that costs nothing in hardware.

Table 3: Complexity analysis of FP-DKLD.

Operations	Σ	×	÷	log	Total
	20	8	0	0	28
Memory	$\lvert P_n\rvert + \psi + 2\beta + L$				

The complexity and storage requirements of a fixed-point DKLD (FP-DKLD) based algorithm utilising a log lookup table and PoT constraint on $|P_n|$ is shown in Table 3. It can be seen that with some simple changes and constraints, the complexity of the anomaly detection algorithm can be greatly reduced—allowing for efficient and high speed hardware implementations. Results obtained from a Xilinx Virtex4 FPGA implementation of the FP-DKLD algorithm are shown in Section 5.

3.4. Information Content Analysis Algorithm. Unlike the KLD-based anomaly detection algorithm just described, the ICA-based method analyses individual input samples rather than aggregate sets of data. The information conveyed by the events is the detection metric utilised. The following is a general description of the algorithm.

First and foremost, it is necessary to establish the type of event that is under observation. This can be any property that is associated with the signal under test (e.g., instantaneous amplitude, phase, or power). If the event type chosen is measurable directly from the signal envelope, demodulation is unnecessary for anomaly detection. The ICA algorithm utilises supervised learning to establish a reference histogram (and hence probability) of events; therefore, some clean signal is required for training. Once the reference histogram is obtained, online analysis of the signal under test can commence. Events from the test signal are extracted and used to update the reference histogram. This yields updated event probabilities and hence the associated information content. If the information content $I_n(x)$ of any event x_n, at time n, is above some predefined threshold I_{th}, an anomaly may be present.

Once again, it is clear that the event histogram plays a central role in the anomaly detection algorithm. It has been stated previously in Section 3.1 and illustrated by Figure 1 that the number of bins utilised, β, has a significant impact on the sensitivity of the histogram and hence the effectiveness of the detection algorithms. When β is too small, anomalous events may not be detected due to poor sensitivity—leading to missed detections. On the other hand, when β is too large, even nominal events will appear to have low probability—leading to a large number of false positives. It is therefore necessary to find a β that offers a good balance between sensitivity and probability of detecting false positives.

The event histograms shown in Figure 1(b) reveal another potential challenge for the ICA algorithm. It can be seen that the histograms have long tails with numerous low probability (i.e., high information content) events even when the signal is behaving nominally. Although this is expected for any analogue signal transmitted over a lossy physical channel, it raises the possibility that numerous false positives are observed at a detector that employs a simple information content threshold. It is certainly possible to reduce the number of tail events by using a smaller number of bins, but that leads to reduction in sensitivity and hence an increase in the probability of missed detections.

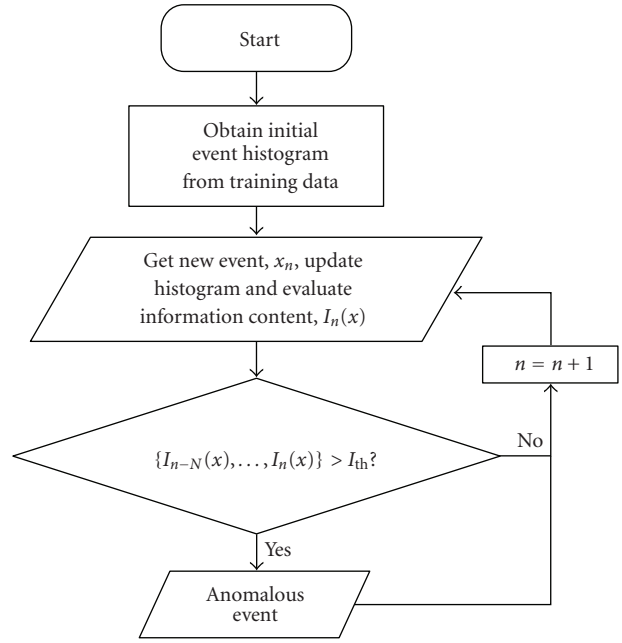


Figure 4: Flowchart of information content analysis algorithm with clustered anomaly detection. First, clean data is used to initialise the reference event histogram and event probabilities. Then, events in the signal under test are used to update the reference histogram and event probabilities. The updated values are used to estimate the information content of the events. If the information content of $N+1$ contiguous events exceeds a predefined threshold, I_{th}, an anomaly may be present in the signal under test.

Examination of the interference scenario in Figure 1(a) reveals an important distinction between anomalous events and the underlying signal—anomalies tend to appear in clusters while nominal low-probability signal events are decidedly "singular." This difference is the key to reducing the number of false positives while still maintaining a low rate of missed detections. The proposed algorithm is easily augmented to benefit from this insight: instead of triggering on individual high information content events, the detector must search for contiguous groups of events that exceed the predefined information content threshold.

A flowchart of the algorithm with simplified clustering is shown in Figure 4. The general approach is as before, with the exception of the last step. With the simple clustering extension, detection of an anomaly is signalled only when a contiguous sequence of N previous events and the current event exceeds a predefined information content threshold. Sequence detection is used rather than full two-dimensional clustering to minimise the complexity of the algorithm. This is permissible since a sequence can be considered as a one-dimensional cluster. The effect of the cluster size utilised on the detector performance is examined in Section 4.2.5.

Table 4: Runtime complexity analysis of ICA.

Operations	Σ	×	÷	log	Total
	2	0	1	1	4
Memory	$\beta + N + 2$				

The discrimination threshold is an important aspect of any detector. While the optimum threshold is problem and cost function specific, it is generally chosen to minimise missed detections while still maintaining a low rate of false positives. For the proposed anomaly detection algorithm, it is not possible to define a single information content threshold, I_{th}, that is suitable for use with any arbitrary signal. I_{th} is signal specific and may be set automatically using information obtained from the clean training data. After the reference event histogram and probabilities have been estimated, the reference information content associated with each event type can be easily computed using (4). For β bins, the standard deviation, $\sigma_{\text{I}\langle\beta\rangle}$, of the reference information content provides a measure of the spread and may be used to obtain the threshold:

$$I_{\text{th}} = m\sigma_{\text{I}\langle\beta\rangle}. \tag{9}$$

m is a multiplicative factor greater than 1. The effect of I_{th} on detector performance is investigated in Section 4.2.3.

Due to the simplicity of the ICA algorithm, its runtime operational complexity and memory requirements are negligibly small. On completion of the initial training phase, a small buffer capable of holding just $\beta + 1$ elements is required to store the event histogram and the total events count. At runtime, analysis of an event requires 2 additions to increment the relevant bin count and the total events count. Division of the incremented bin count by the total is then needed to obtain the event probability. After the probability is computed, a single base-2 logarithm is needed to calculate the event's information content. An additional buffer capable of holding $N + 1$ elements is also needed to accommodate information content clustering. A summery of the complexity analysis is provided in Table 4. It reveals that in addition to being negligibly small, the fixed runtime operational complexity is independent of any algorithmic parameter (e.g., histogram resolution)—suggesting that fast and efficient implementations for power limited hand-held devices are possible.

4. Results

In order to evaluate the performance of the proposed anomaly detection schemes, signals with different classes of abnormalities are employed as test cases. All of the signals under test are actual radio frequency transmissions captured using spectrum analysis hardware and therefore represent scenarios likely to be encountered by real world wireless devices.

Analyses of the test signals are provided in the following section while detailed performance analyses of the algorithms based on parameters such as histogram bin resolution, data window size, sampling rate and cluster length are provided in Section 4.2.

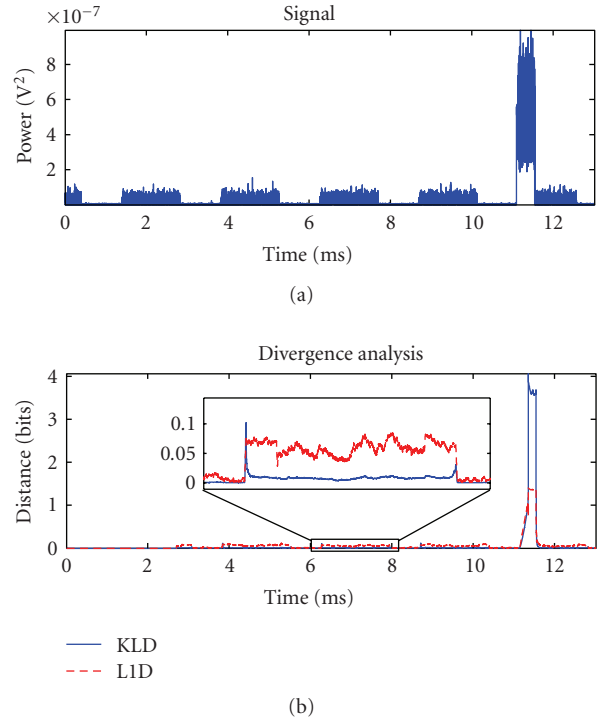


Figure 5: (a) WLAN signal with interference at 11.5 ms from a Bluetooth device. (b) Both KLD and L1D analyses of the signal result in detection of the anomaly. KLD appears to be better than L1D since it leads to a much larger peak and lower noise.

4.1. Data Analysis. Of the four data sets available, the first three are used to demonstrate the PMF divergence analysis (KLD/L1D)-based technique while the last is used to demonstrate the ICA-based technique.

4.1.1. Test Signal A. The signal is shown as a time series in Figure 5(a). It consists of a single WLAN frame that repeats with a period of 2.45 ms and a burst of interference from a Bluetooth device that is visible at 11.5 ms. The signal is similar to that shown earlier in Figure 1.

Both KLD and L1D are used to analyse the signal for the purpose of obtaining results that can be directly compared. Two windows with a duration of 256 μs each are employed to process the time series signal. The window centres are separated by 2.45 ms to match the WLAN frame repetition interval. The windows estimate the PMFs of the signal power. The number of histogram bins utilised is the optimal value (51 in this case) as obtained from the algorithm proposed by Shimazaki and Shinomoto [1]. In any case, it is shown in a subsequent Section (4.2.4) that the number of histogram bins used does not have a significant impact on the outcome—therefore, an arbitrary but reasonable choice such as 32 can also be used instead.

The result of the analysis is also shown in Figure 5. Both KLD-and L1D-based methods are successful at detecting

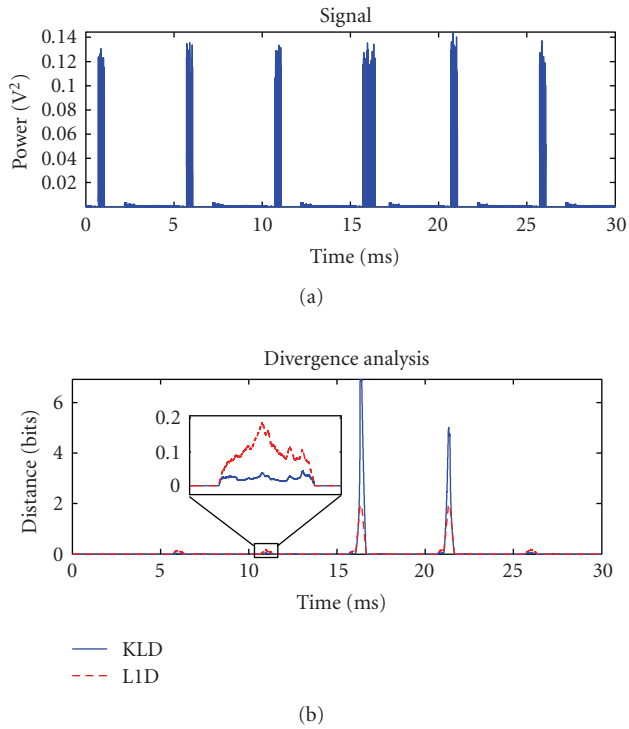


Figure 6: (a) Wireless broadband (WiBro) signal. The uplink subframe at 16 ms has a longer duration than others and is therefore unusual in this context. (b) KLD and L1D analysis both reveal the anomalous segment of the WiBro signal. A second peak is obtained when the signal returns to normal—this is due to the twin-windowing nature of the anomaly detection algorithm. Once again, KLD analysis results in a larger peak and lower noise.

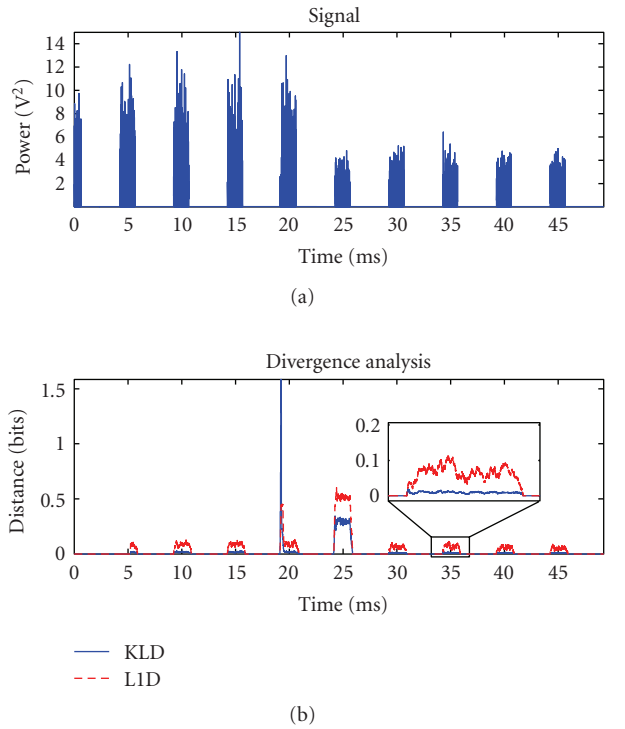


Figure 7: (a) Message exchange between a mobile terminal and a base station using the WiBro communications protocol. An extra command sequence in the uplink subframe at 20 ms initiates the power control loop. (b) The extra command sequence (20 ms) and subsequent power change (25 ms) are both revealed by KLD/L1D analyses of the signal. A larger KLD peak results from the anomalous command but not from the power change. L1D also leads to larger noise levels compared to KLD.

the presence of the anomaly (BT interferer). However, it is clear that KLD is the better choice as it produces a larger peak compared to L1D when the anomaly is detected. The baseline noise level with KLD is also much lower than that obtained with L1D—confirming the hypothesis presented in Section 3.

4.1.2. Test Signal B. The second signal under test is a wireless broadband (WiBro) signal. It is shown in Figure 6(a). Due to the proximity of the recording equipment to the mobile terminal (MT), the uplink (UL) subframes show a higher power level then downlink (DL) subframes. From the plot, it can be seen that the UL subframe at 16 ms is longer than any of the other UL subframes. In context of this particular signal snapshot, this behaviour is unusual and hence can be considered to be anomalous. Once again, two windows with a duration of 256 μs each are employed to estimate the signal power PMFs. The window centres are separated by 5 ms—corresponding to the frame period of the signal. The optimal histogram bin allocation scheme in [1] is once again used to determine the number of bins utilised ($\beta = 51$).

Analysis of the signal is also shown in Figure 6. A sharp peak in the divergence at 16 ms reveals the presence of the unusual UL subframe. A second peak is obtained when the signal returns to normal in the following UL subframe. Once again, the superiority of KLD over L1D as a divergence metric is demonstrated by the larger peaks and lower baseline noise levels.

4.1.3. Test Signal C. The third signal used to test the detection capabilities of the divergence-based algorithm is shown in Figure 7. It depicts communication between a mobile terminal and base station using the WiBro standard. Since the recording is made at the MT, there is significantly more power in the UL subframes. Although unnoticeable in the time series, the UL subframe at 20 ms contains an additional command sequence that triggers the subsequent change in the transmit power observed at 25 ms. As a result, there are effectively two unusual events in the signal: the extra command and the subsequent change in power level. The parameters utilised for analysis of the signal are identical to those used in Section 4.1.2.

The divergence analysis plot in Figure 7 shows that both anomalies can be successfully detected using KLD and L1D. Since the width of a KLD peak corresponds to the temporal duration of the anomaly responsible, the first peak at 20 ms is very sharp as it is due to the extra command sequence in the UL subframe. Since the subsequent change in power at 25 ms affects the entire UL subframe, the second peak is much broader and spans the entire subframe.

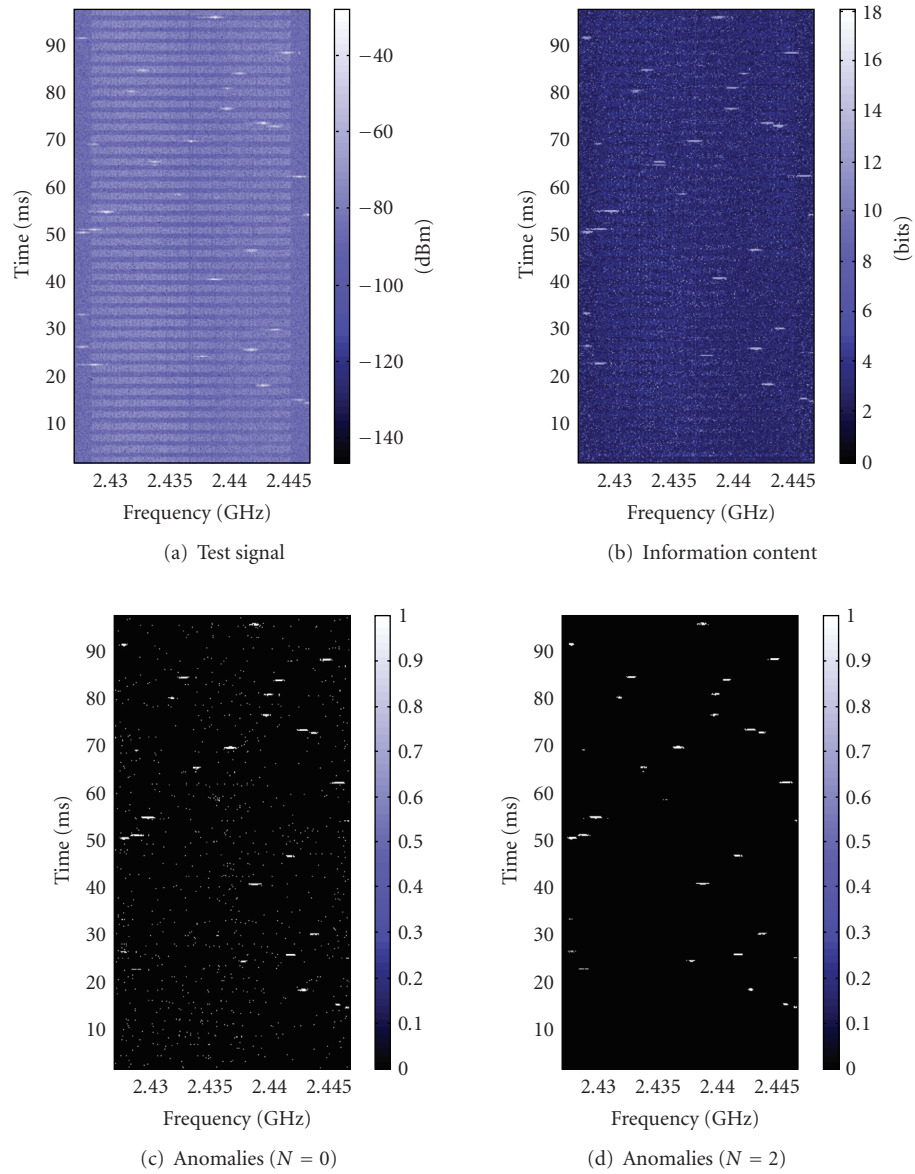


Figure 8: Analysis of a WLAN signal with interference from a Bluetooth device. (a) Spectrogram of the original signal. Data bursts from the Bluetooth device are clearly visible as high power, lightly shaded patches. (b) Information content of events (power density at any given time-frequency point) in the signal. (c) Anomalies detected (light patches) using a threshold of $1.25\sigma_{\mathrm{I}\langle 16\rangle}$ (7.15 bits) and a cluster length of 1. The result is noisy and there are a lot of false positives (appearing as singular, lightly shaded spots). (d) Anomalies detected when the threshold is left unchanged at $1.25\sigma_{\mathrm{I}\langle 16\rangle}$ and the cluster length is increased to 3. The outcome is now much cleaner with virtually zero false positives.

The plot also reveals that for the second anomalous event (power change), KLD is smaller than the associated L1D. This is not unexpected since it has been hypothesised in Section 3 that for differences that lead to an L1D of $2\ln 2$ or smaller, L1D can be larger than KLD. It is also the reason why L1D generally leads to larger baseline noise levels compared to KLD.

4.1.4. Test Signal D. The final test signal is used to evaluate the ICA algorithm. It is similar in nature to the signal shown in Figure 1(a). Spectrogram of the test signal is shown in Figure 8(a). It is a much longer signal with numerous interference events to provide a statistically significant sample size. The plot depicts a real WLAN signal with Bluetooth interference captured over the air-interface. The WLAN signal consists of a single frame that is repeated periodically by a vector signal generator. The characteristic frequency hopping pattern of the Bluetooth device marks the locations of the interference (anomalous) events.

The signal spectrogram is estimated from the time series using nonoverlapping Hamming windows that are $64\,\mu$s long. A 1024 point FFT (fast Fourier transform) is used to obtain a frequency resolution of approximately 20 kHz. The signal event under observation is the instantaneous power

density at any given time-frequency coordinate. The first 10 ms of the signal is assumed to be free from interference and is therefore used for training purposes. 16 equally spaced histogram bins divide the range between the maximum and minimum power densities observed in the training data. The $\sigma_{I\langle 16\rangle}$ for the training data is 5.72 bits.

Figure 8(b) shows the information content of events in the test signal. As expected, the anomalous events have a higher information content and they are highlighted while the regular underlying structure is suppressed. The plot also shows that there is a lot of noise (tiny spots of high information content) from individual low-probability signal events that are otherwise nominal. The reason for this behaviour has been outlined in Section 3.4.

Anomalies detected using a threshold of $1.25\sigma_{I\langle 16\rangle}$ and a cluster length of 1 (i.e., only the current event) are shown in Figure 8(c). It is immediately obvious from the large number of small, lightly shaded spots that there are a lot of false positives. Again, singular low-probability signal events are responsible since they can potentially have higher information content than actual anomalous events. Keeping I_{th} the same and increasing the cluster length to 3 yields the result shown in Figure 8(d). It reveals that a simple change in the cluster length is sufficient for reducing the number of false positives to virtually zero.

4.2. Performance Analysis. The analyses of test signals presented in Section 4.1 show that both algorithms perform well for the parameter combinations chosen. In order to investigate and quantify the impact of other parameter choices, it is necessary to define and utilise metrics that reflect performance.

For the divergence-based technique, the ratio between the anomaly detection peak and the maximum of the baseline noise level is a good indicator of performance since it is a reflection of the range over which a threshold can be applied. It can be seen from the results presented in Figures 5, 6, and 7 that KLD is an extremely effective discriminator for statistical changes in the observed data. Even with such *real test vectors captured over-the-air*, the KLD peaks produced by anomalous events are many orders of magnitude larger than baseline noise levels associated with nominal data. As a result, 100% probability of detection can be achieved over a wide range of KLD threshold values (the anomalous peak is approximately 140 times as large as the background noise level in Figure 6) while still guaranteeing a 0% probability of false positives—making such classical measures of detector performance inadequate for gauging the true extent of the algorithm's performance.

Another reason against the suitability of classical performance measures such as receiver operating characteristic (ROC) curves is the scarcity of available test data. Probability of detection and false positives are inherently statistical measures of performance that require a large sample size to produce meaningful results. Since the focus of this work is exclusively on practical applications of the proposed algorithm, the number of test vectors available is limited and each test signal (A, B and C) contains only 1 or 2 anomalous events. So instead of attempting to extract questionable probability measures from the limited data set, a measure of the difference between the height of the anomalous peak and the baseline noise level is utilised to quantify the observed performance.

When KLD is used as the measure of divergence, the KLD ratio (KLDR) is defined as

$$\mathrm{KLDR} = \frac{\mathrm{KLD}_{\mathrm{anom}}}{\mathrm{KLD}_{\mathrm{bg}}}, \tag{10}$$

where $\mathrm{KLD}_{\mathrm{anom}}$ is the maximum of the detection peak and $\mathrm{KLD}_{\mathrm{bg}}$ is the maximum of the background baseline noise level. KLDR is the metric that is used to quantify the algorithm's performance.

For the ICA-based algorithm, the circumstances are different. The test set (signal D) contains a sufficient amount of nominal and anomalous events to allow the use of more traditional performance metrics. Performance is measured in terms of the detector true positive rate (R_{tp}) and false discovery rate (R_{fd}). R_{tp} is defined as the ratio of the number of correctly detected anomalous events (Σ_{tp}) to the total number of anomalous events present (Σ_{ta}):

$$R_{\mathrm{tp}} = \frac{\Sigma_{\mathrm{tp}}}{\Sigma_{\mathrm{ta}}} = 1 - \frac{\Sigma_{\mathrm{md}}}{\Sigma_{\mathrm{ta}}}, \tag{11}$$

where Σ_{md} is the number of anomalous events that missed detection. R_{fd} is the ratio of false positives (Σ_{fp}) to the total number of anomalies detected (includes both Σ_{fp} and Σ_{tp}) [29]:

$$R_{\mathrm{fd}} = \frac{\Sigma_{\mathrm{fp}}}{\Sigma_{\mathrm{fp}} + \Sigma_{\mathrm{tp}}}. \tag{12}$$

R_{fd} is preferred over the more common false positive rate (R_{fp}) as it is more useful in this context. R_{fp} is defined as the ratio between Σ_{fp} and all nonanomalous events (Σ_{tn}) in the signal:

$$R_{\mathrm{fp}} = \frac{\Sigma_{\mathrm{fp}}}{\Sigma_{\mathrm{tn}}}. \tag{13}$$

It is also known as the false alarm rate. Since Σ_{tn} is a very large number, R_{fp} is close to zero for most parameter combinations and therefore does not adequately reflect the variations observed in detector performance.

To summarise, KLDR is used to evaluate the performance of the KLD-based algorithm while R_{tp} and R_{fd} are used to evaluate the ICA-based algorithm.

4.2.1. Sampling Rate. Continuous processes such as time-series must be sampled before the anomaly detection algorithms can be applied. The sampling frequency employed is crucial as it dictates the size of the input data sets, $|P_n|$, and therefore the memory utilisation of the KLD-based algorithm—as indicated in Table 2. For a given window length, a higher frequency implies that more data samples have to be stored and sorted to construct the histograms. If the frequency is too low, small scale signal features and anomalies may be lost. According to the Nyquist sampling

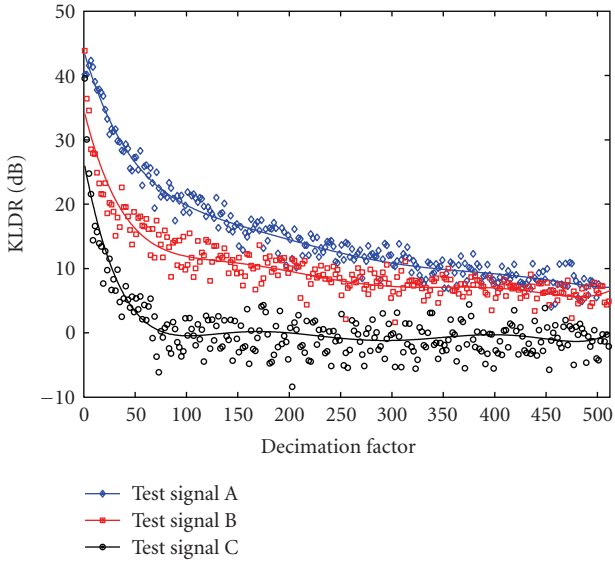


Figure 9: Performance of the KLD-based anomaly detection scheme under various data sampling rates. Decimation factor refers to the amount by which the input signal is undersampled relative to the signal bandwidth. A factor of unity corresponds to a sampling frequency equal to the signal bandwidth. For two of the test cases (A and B), a KLDR well above 10 dB can be maintained for a window length of 256 μs, histogram bin count of 32, and a decimation factor of 100. Signal C is unable to accommodate such high decimation rates.

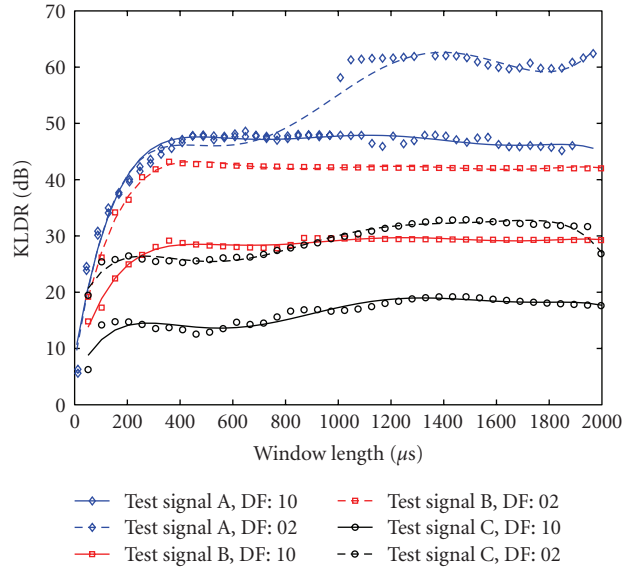


Figure 10: Performance of the KLD-based anomaly detection scheme under various window lengths and hence the input sample size $|P_n|$. Decimation factors (DF) of 2 and 10 are utilised and the number of histogram bins is 32. In all cases, an increase in the amount of data leads to a better performance. However, beyond a certain window length, the performance is no longer strongly affected.

criterion, a signal must be sampled with a frequency at least twice as large as its bandwidth to be reconstructible. For wideband signals this leads to a very high sampling frequency and hence a prohibitively large volume of data—heavily increasing the resource requirements of the proposed scheme. Since neither of the proposed algorithms require the time-series to be reconstructible, a far lower sampling frequency can be used instead. Figure 9 shows how the performance of the KLD-based anomaly detection scheme is affected by undersampling of the input time series. The window length utilised is 256 μs and the histograms used to construct the PMFs are 32 bins wide. The amount by which the input time-series is undersampled relative to the bandwidth is defined as the decimation factor. Therefore, a factor of unity implies that the signal is sampled at the same frequency as the signal bandwidth.

The results indicate that decimation factors as large as 500 can be successfully employed depending on the type and duration of the anomaly present. For test signals A and B, a KLDR of more than 10 dB can be maintained even with a decimation factor of 100. This is an important result as it indicates that satisfactory performance levels can be maintained with little input data and hence memory-limited implementations of the algorithm. At high decimation factors, performance is poor for test signal C. This is because the first anomaly (extra command sequence) is temporally brief and is likely to be lost when the signal is heavily undersampled. As for the second anomaly in the signal, the change in power is simply not large enough to produce a significant increase in the divergence.

Results for the ICA-based method are not shown since the input sampling rate has no bearing on the complexity or memory requirements of the algorithm—as shown in Table 4. The only requirement then on the sampling rate is that it must be fast enough to capture events that are suspected of being anomalous.

4.2.2. Window Length. The window size, and hence the input data set size $|P_n|$, is another parameter that is relevant for the KLD-based algorithm but not the ICA-based algorithm. The effect of the PMF estimation window size on the performance of the algorithm shown in Figure 10. The number of histogram bins utilised is 32. Results are shown for undersampling factors of 2 and 10. At lower decimation factors, more data is available and the KLDR improves uniformly across all window sizes for signals A and B. At smaller window sizes, performance for signal A is unaffected by the choice of the decimation factor due to the relatively long duration of the anomaly. This is because even at a decimation factor of 10, a sufficient number of anomalous samples are represented in the PMF.

As anticipated, the performance is poor at small window sizes where the amount of data available is insufficient to adequately model the underlying PMFs. Increasing the window length leads to an improvement of the performance. However, for signals B and C, the gains become marginal for windows larger than approximately 400 μs. The transition shown by signal A at a window length of 1 ms for a decimation factor of 2 is due to a sudden reduction in the KLD noise at the frame edges (as seen in the zoomed-in

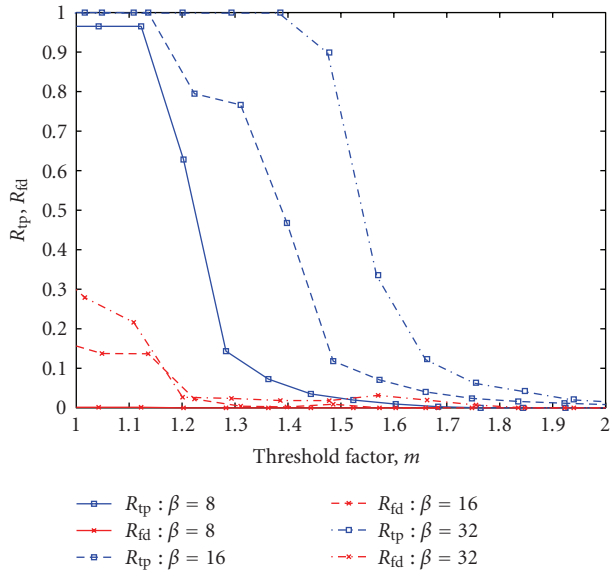


FIGURE 11: Effect of information threshold (9) on detector R_{tp} and R_{fd} for the ICA algorithm. The cluster length utilised is 3. Histogram bin sizes of 8, 16, and 32 are used for comparison. R_{tp} for each β shows a sharp decrease when I_{th} (i.e., $m \cdot \sigma_{I\langle\beta\rangle}$) is increased above a certain limit. Majority of the anomalous events have an information content less than this and miss detection. $m = 1.35$ and $\beta = 32$ yields the best performance.

segment of Figure 5) while the detection peak remains at approximately the same level. It is no coincidence that the duration of the interframe spacing for the signal is also 1 ms. It is a signal feature that is detected by the algorithm alongside the actual anomalous events. When the window size is increased beyond this feature size, it can no longer be resolved effectively by the detection algorithm and lead to a decrease in the noise level seen at the frame edges.

The initial KLDR improvements with increasing window size are due to improvements in the PMF estimates which in turn lead to a reduction in the baseline KLD levels. At larger window sizes, the anomalous samples represent smaller fractions of the data and hence contribute less to the shape of the estimated PMF—resulting in a decrease of the KLD due to the anomaly. As the background levels are also reduced by an increase in the data size, the overall KLD ratio (i.e., the KLDR) remains relatively constant.

4.2.3. Information Threshold. As stated previously in Section 3, the discrimination threshold (9) is an important aspect of any detector. The impact of I_{th} on R_{tp} and R_{fd} of the ICA-based algorithm is investigated using a cluster length of 3 and histogram bin sizes of 8, 16, and 32. The result of the analysis is shown in Figure 11.

The plot shows that there is a hard I_{th} boundary for each β after which R_{tp} drops rapidly. This implies that the majority of the anomalous events share similar characteristics and convey information equivalent to that boundary. When I_{th} is increased further through the use of a larger threshold factor m, R_{tp} approaches zero due to an ever increasing number of missed detections.

At low-information content thresholds, R_{fd} is also high—specially for high values of β. As explained earlier in Section 3.1, a higher resolution makes the detector more susceptible to noise, leading to an increase in the number of false positives and hence the R_{fd}.

The impact of D_{th} on the performance of the KLD-based method has not been investigated and therefore cannot be shown. The reasons for this are as follows.

(i) The number of anomalous events available is insufficient to investigate statistical trends.

(ii) Detection is often guaranteed for a wide range of thresholds due to the large KLDR (greater than 30 dB) values that are observed.

It is the second reason that generally makes it straightforward to choose a suitable D_{th}.

A mathematical treatment of the impact of D_{th} and I_{th} on detector performance is beyond the scope of this article. Such a framework requires well-defined theoretical models of the data distribution which are difficult to obtain for real data vectors. Equations derived using the simplifying assumption that the distributions belong to a well-known class such as Gaussian would be of little use in context of the test signals used in this paper. Since the signals do not conform to any standard probability density function, it is out of necessity that the thresholds are determined empirically.

4.2.4. Histogram Resolution. Histogram bin resolution, represented by the parameter β, is of relevance to both of the proposed algorithms. Figure 12 shows how performance of the KLD-based algorithm is affected by the choice of the number of histogram bins used to classify the input data and estimate the PMFs. The window length is set at 256 μs and results are shown for decimation factors of 2 and 10. Once again, the smaller decimation factor provides uniformly improved performance over the entire range of β values. The only exception is signal A where the performance for smaller β values appears to be independent of the decimation factor used. The reasons for this is the relatively long duration of the anomaly—as explained previously in Section 4.2.2.

The only trend common to all three signals is that the performance changes little with increasing bin numbers, with signal C showing an optimum in the vicinity of $\beta = 55$. This indicates that the behaviour observed is specific to the type of anomaly present in a signal. While the number of bins utilised does not appear to have a significant impact on the performance of the scheme for a fixed amount of data, the decrease observed is due to noisier PMF estimates that are obtained for larger values of β. Noisy PMFs lead to larger background KLD values and hence a reduced KLDR.

Figure 13 shows how the R_{tp} and R_{fd} vary for the test signal (Figure 8(a)) with the number of histogram bins utilised. The cluster length utilised is 3 and I_{th} of $1.2\sigma_{I\langle\beta\rangle}$, $1.4\sigma_{I\langle\beta\rangle}$, and $1.6\sigma_{I\langle\beta\rangle}$ are used for comparison. The plot reveals that when $\beta = 4$, R_{tp} is zero and R_{fd} is unity for all thresholds tested. This is because the sensitivity is very low and no anomalies can be detected ($R_{tp} = 0$). Events exceeding the threshold are low-probability signal events and

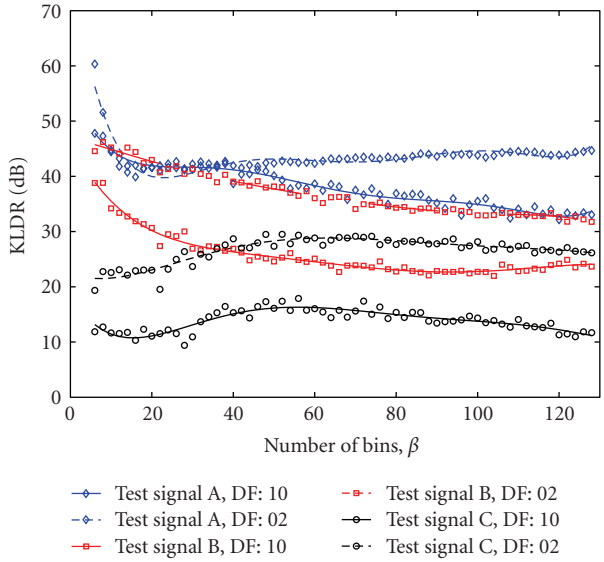


FIGURE 12: Performance of the KLD-based anomaly detection scheme under various histogram bin counts. The window size is set at 256 μs and the decimation factors used are 2 and 10. Generally, a larger number of bins lead to poorer performance due to increased noise in the estimates. However, the rate of change is small and therefore the drop in performance is insignificant over a wide range of bin resolutions.

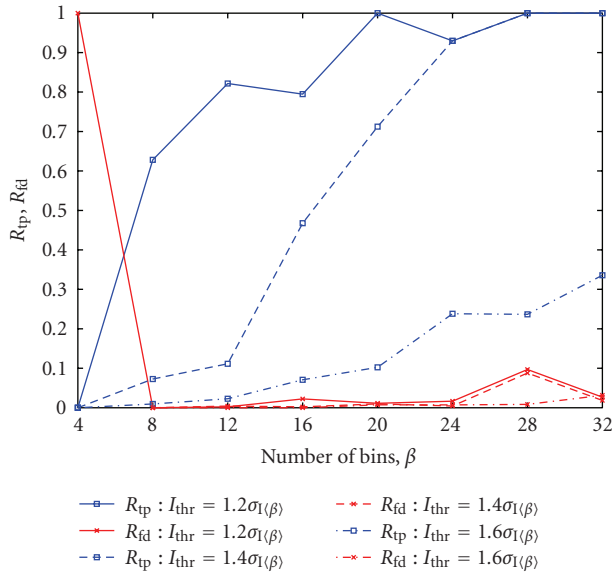


FIGURE 13: Effect of histogram resolution on detector R_{tp} and R_{fd}. The cluster length is 3 and thresholds are $1.2\sigma_{I\langle\beta\rangle}$, $1.4\sigma_{I\langle\beta\rangle}$, and $1.6\sigma_{I\langle\beta\rangle}$. R_{tp} improves with resolution while R_{fd} deteriorates. $\beta = 20$ with a threshold of $1.2\sigma_{I\langle 20\rangle}$ yields the best performance.

hence are all false positives ($R_{fd} = 1$). As β is doubled to 8, the resolution improves and there is a corresponding increase in the R_{tp}. The R_{fd} also drops to a negligibly small value. As β is increased further, the R_{tp} increases due to better detector resolution. The R_{tp} improvements come at a cost, the detector is more susceptible to noise at higher

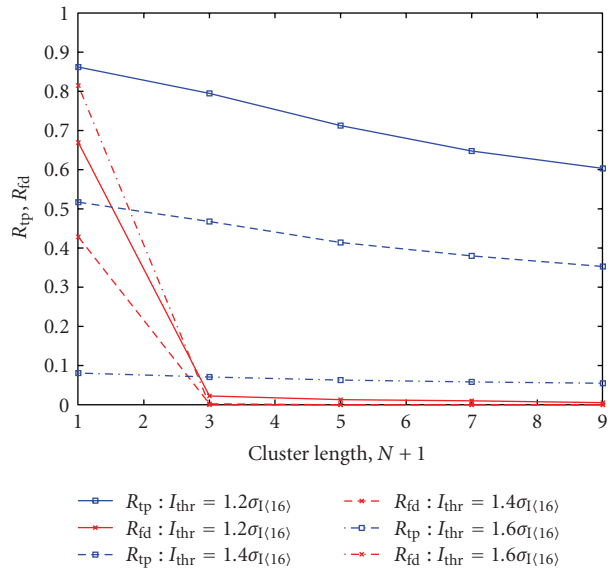


FIGURE 14: Impact of cluster length on detectors R_{tp} and R_{fd}. The number of histogram bins used is 16 and the thresholds utilised are $1.2\sigma_{I\langle 16\rangle}$, $1.4\sigma_{I\langle 16\rangle}$, and $1.6\sigma_{I\langle 16\rangle}$. Even the smallest cluster length ($N = 2$) is shown to provide a significant improvement in the R_{fd}.

resolutions. This is evident from the gradual increase in the R_{fd}.

Comparison between the three detection thresholds reveals that a higher R_{tp} is achieved with a lower threshold. Unfortunately, this also leads to a higher R_{fd}. This behaviour is in accordance with the explanation provided in Section 4.2.3.

4.2.5. Cluster Size. In order to investigate the impact of the information cluster length, N, on the ICA-based detector, β is set at 16 and the analysis is performed for I_{th} of $1.2\sigma_{I\langle 16\rangle}$, $1.4\sigma_{I\langle 16\rangle}$, and $1.6\sigma_{I\langle 16\rangle}$ on the test signal shown in Figure 8. The result of the analysis is shown in Figure 14.

The significance of clustered anomaly detection is immediately obvious. With $N = 0$, when clustering is not performed, there are an overwhelming number of false positives. This is indicated by the high R_{fd}. As soon as clustering is applied by setting $N = 2$, a dramatic drop in the R_{fd} is observed—showing that even minimal anomaly clustering is sufficient to yield a massive improvement in detector performance. By lowering the R_{fd}, clustering also allows a lower I_{th} to be used to achieve a higher R_{tp}.

The impact of anomaly clustering on R_{tp} for a given I_{th} is relatively low. As cluster size is increased, a gradual decrease is observed in the R_{tp}. This is expected since larger cluster sizes lead to missed detections around the edges of the interference patterns. The plot also shows that higher thresholds lead to lower R_{tp} for a given cluster size. This is also expected since a higher information content threshold leads to a higher number of missed detections.

From the analysis performed on the test signals, it is clear that it is challenging to determine a set of parameters that are inherently optimal for the anomaly detection algorithms

proposed. This is due to the fact that the optimal parameter set depends on a number of problems-specific factors such as the duration of the anomaly and the dynamic range of the signal. It may be possible to develop adaptive variants of the algorithms that automatically find the best parameter combinations subject to some performance criterion but that is beyond the scope of this paper.

The ICA-based algorithm is particularly sensitive to the parameters utilised. Generally, it is seen that parameter values that increase the R_{tp} (good) often also lead to an increase in the R_{fd} (bad) and vice versa. Trade-offs must therefore be made to meet the required detector performance characteristic (low R_{fd}, moderate R_{tp} or high R_{tp}, moderate R_{fd}). A moderate number of bins ($\beta = 20$), small cluster size ($N = 2$), and a threshold of $1.2\sigma_{I\langle 20\rangle}$ bits ($m = 1.2$) provide a good balance between R_{tp} (1.0) and R_{fd} (0.01) for this particular test vector (signal D).

The KLD-based algorithm on the other hand is much more robust with respect to the parameter combinations utilised. The results clearly show that performance better than 30 dB of KLDR can be easily obtained with reasonable choice of parameter values such as a window length of 256μs, $\beta = 32$ and decimation factor of 2.

5. Hardware Platform

The FP-DKLD version of the detection algorithm presented in Section 3.3 has been implemented on a Xilinx Virtex-4 ML402 SX XtremeDSP Evaluation Platform to serve as a proof of concept and allow the testing of signals in real time. To facilitate and accelerate code development, Xilinx SystemGenerator 10.1 is used in conjunction with MATLAB R2007a for the primary design flow. The implemented design runs at a clock speed of 80 MHz and is capable of processing input with a 10 MHz sample rate. The hardware chain used to test and validate the FP-DKLD implementation is shown in Figure 15.

The Agilent E4438C ESG signal generator simultaneously provides analogue and digital versions of the signal under test. The digital data stream is connected to the FPGA platform via the Agilent N5102A Digital Signal Interface Module (DSIM) while the analogue signal is connected to a oscilloscope for display. The DSIM conditions the data (word size, bit alignment, clock relationship settings) and provides a synchronous clock signal that is used to drive the FPGA core. The trigger output from the FPGA platform is also connected to the oscilloscope via a digital probe so that it can be directly compared against the signal under test.

A pair of Wireless Broadband (WiBro) signals known to contain a number of different anomalous data segments are used to test the hardware platform. The design is configured with $|P_n| = 4096$ (320 μs) and $\beta = 8$. It is not necessary to down sample the input data stream since the implemented design is capable of processing the input at its original rate.

The DSIM module provides the samples to the FPGA as 12-bit words in 2s complement format. The sample and DSIM clocks are set at 10 MHz and 40 MHz respectively-providing 4 clock cycles per input sample (CCPS). Although

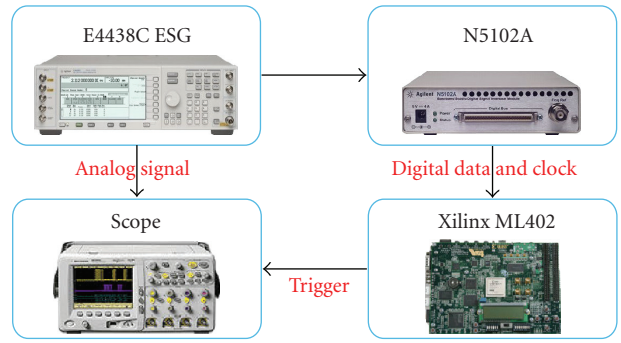


Figure 15: Block diagram of the hardware test-bed. The E4438C ESG signal generator produces the signal under test in both analogue and digital formats. The digital signal is passed to the FPGA platform via the N5102A digital signal interface module while the analogue signal is fed into the oscilloscope. The trigger signal from the FPGA core is also connected to the oscilloscope via a digital probe for comparison.

the design requires 8 CCPS, the DSIM is only capable of providing a maximum of 4 CCPS. To obtain the required 8 CCPS, the clock signal is doubled on the FPGA using an on-chip digital clock manager (DCM) module. Use of a DCM also has added benefit of providing clock buffering and deskewing.

5.1. Test Signal I. Figure 16 shows the result of analysing the first WiBro signal using the FP-DKLD implementation of the algorithm. The signal analysed is identical to that shown in Figure 6 and analysed in Section 4.1.2. It is clear from the oscilloscope trace that one of the UL frames is longer than the others and hence is anomalous. With $D_{th} = 0.0313$, the FPGA implementation of the algorithm clearly succeeds in detecting the signal anomaly. The first trigger event obtained (A) coincides exactly with the anomalous segment of the unusual UL frame. A second trigger event (B) is observed when the UL frame structure subsequently returns to normal and the anomalous segment is no longer present.

5.2. Test Signal II. The second WiBro signal tested is shown in Figure 17. It is identical to that shown in Figure 7 and analysed in Section 4.1.3 with the exception of an additional change in the timing structure. Analysing the signal with $D_{th} = 0.0625$ is seen to produce five trigger events—corresponding to the three anomalous conditions known to be present in the signal.

Trigger events A and B are due to a momentary disruption in the natural frame period of the signal. The first event marks the position where the UL frame should have been but is not while the second event marks the opposite: finding a UL frame where there should be none.

Events C and D are caused by a very brief command sequence at the beginning of the fifth UL frame that causes the power control loop to be initiated—which is then responsible for event E. Although invisible to the naked eye, the algorithm succeeds in locating the anomalous command sequence as clearly demonstrated by trigger event

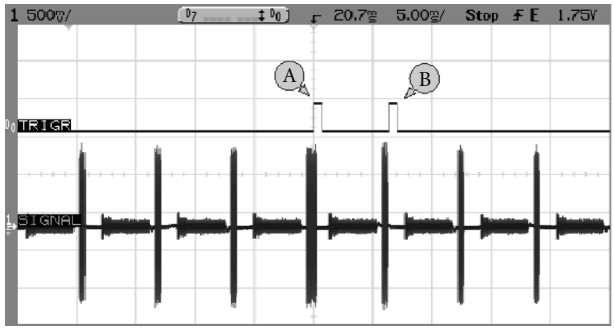


Figure 16: Oscilloscope trace of a WiBro signal with a single anomalous frame and the associated trigger events. The KLD threshold is 0.0313. Trigger event A marks the start of the unusual segment of the anomalous frame. A second trigger event, B, is also obtained in the subsequent frame due to the disappearance of the anomalous feature.

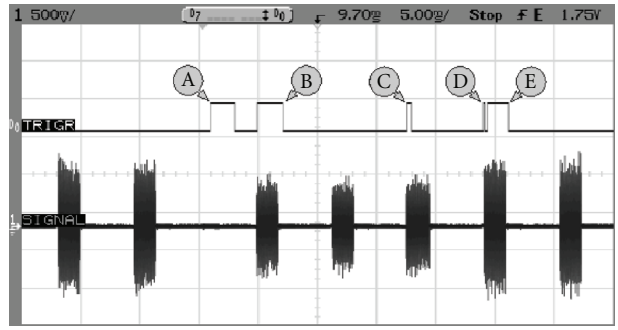


Figure 17: Oscilloscope trace of a WiBro signal with multiple anomalous events and the associated trigger events. The KLD threshold is 0.0625. Triggers A and B are caused by a momentary change in the signal period. Trigger C flags the presence of a very brief command sequence that leads to a signal power level change. Trigger D marks the position in the subsequent frame where the power change inducing command was previously present and finally, trigger E corresponds to a sudden change in the overall signal power level.

C. Since that command sequence is no longer present in the subsequent UL frame, its disappearance is marked by trigger event D.

Once initiated, the power control loop causes a sudden increase in the signal power level. This behaviour can be considered to be anomalous and is flagged by trigger event E. There are no other events associated with the change in power level as the signal power is seen to remain high beyond this point.

6. Summary and Conclusions

Two complementary anomaly detection algorithms utilising information theoretic measures have been presented. Both algorithms are simple to implement and require little a priori information regarding the signal under test. Demodulation of the signal is also not required since the algorithms are capable of processing the baseband signal envelope itself in real time. The information content analysis based method is capable of detecting singular anomalous events while the Kullback-Leibler divergence based method is also able to detect otherwise nominal events that are anomalous purely due to context (e.g., misaligned signal frames). In order to provide this context aware detection of anomalies, the KLD-based algorithm requires the input signal to be periodic.

Analyses of a number of test signals captured over the air show that the KLD-based scheme is successful at detecting all anomalies known to be present. Extensive tests using a software implementation of the algorithm demonstrate that it is robust with respect to parameter choices since satisfactory performance can be maintained with reasonable parameter values even when the input is severely undersampled. With PMF estimation window sizes of 256 μs, 32 histogram bins and factor of 10 undersampling, KLDR of 25 dB or better can be achieved depending on the anomaly present.

Although the primary purpose of the KLD-based algorithm is to act as an anomaly detector, it can also be used to detect frame boundaries in a signal. The modification required is trivial: eliminate the spacing that normally separates the two PMF estimation windows. Since frame boundary detection is expected to reveal the underlying cyclic structure of a periodic signal, it may be used as a precursor to the actual anomaly detection algorithm to automatically learn the period of the signal—thus eliminating the need for any a priori information regarding a test vector.

A variation on the anomaly clustering technique presented in context of the ICA-based algorithm may also be applied to the KLD-based algorithm to further improve detection of anomalous events. Anomalies generally lead to KLD peaks that increase monotonically until some maximum divergence is reached. It may be possible to exploit this observation to improve detection under low SNR conditions by restricting detection to signal segments that lead to monotonically increasing KLD values that are also above some predefined KLD threshold.

Both boundary detection and monotonic sequence detection are techniques that add significantly to the KLD-based anomaly detection algorithm. Therefore, they will be the primary focus of work done in the future on this subject.

In addition to the MATLAB based software, the algorithm has been implemented on a Xilinx Virtex4 FPGA based hardware platform for evaluation under real world physical conditions. The design is highly efficient and capable of processing 10 MHz input signals without requiring any undersampling. Successful tests with a set of WiBro signals indicate that the algorithm is indeed capable of processing high speed test vectors in real time.

Unlike the KLD-based method, the algorithm utilising ICA for anomaly detection does not require the input signal to be periodic. The only piece of information that is needed in advance is the set of reference event probabilities. A training data set known to be clean can be used to obtain the reference probabilities prior to analysis. The complexity and memory requirements of the algorithm are also very low.

It is clear from tests carried out using a software implementation of the algorithm that performance of the

system is strongly affected by the choice of parameters such as histogram resolution, threshold and cluster size. Impact of these parameters on the detector performance by means of the true positive rate and false discovery rate has been analysed and guidelines for appropriate values have been provided. It is shown that a true positive rate of 100% and false discovery rate of 1%—guaranteeing zero missed detections with very few false positives—is possible for the signal tested with a suitable set of parameter choices.

The ICA-based algorithm presented in this paper utilises a histogram with infinite memory, that is, it maintains a record of all samples analysed. Clearly this implies that the information content of anomalous events such as an interference drops over time if they happen with sufficient frequency. If such behaviour is undesirable, it is necessary to implement a windowed histogram. Along with tests against other types of anomalous signals, it will be the focus of further research on the ICA algorithm.

Acknowledgments

The authors would like to thank Roy Macnaughton and Peter Cain for the provision of test data. They greatly acknowledge the financial support of this project from Agilent Technologies and the University Relations Program. Harald Haas acknowledges the Scottish Funding Council support of his position within the Edinburgh Research Partnership in Engineering and Mathematics between the University of Edinburgh and Heriot Watt University.

References

[1] H. Shimazaki and S. Shinomoto, "A method for selecting the bin size of a time histogram," *Neural Computation*, vol. 19, no. 6, pp. 1503–1527, 2007.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[3] S. Haykin, D. J. Thomson, and J. H. Reed, "Spectrum sensing for cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 849–877, 2009.

[4] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 772–776, IEEE, Pacific Grove, Calif, USA, November 2004.

[5] D. Cabric, A. Tkachenko, and R. W. Brodersen, "Spectrum sensing measurements of pilot, energy, and collaborative detection," in *Proceedings of the IEEE Military Communications Conference (MILCOM '06)*, pp. 1–7, Washington, DC, USA, October 2006.

[6] M. Öner and F. Jondral, "Air interface recognition for a software radio system exploiting cyclostationarity," in *Proceedings of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '04)*, vol. 3, pp. 1947–1951, Barcelona, Spain, 2004.

[7] C. Krügel, T. Toth, and E. Kirda, "Service specific anomaly detection for network intrusion detection," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 201–208, ACM, Madrid, Spain, March 2002.

[8] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: finding unusual medical time series," in *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pp. 329–334, IEEE, Dublin, Ireland, June 2005.

[9] R. J. Bolton and D. J. Hand, "Statistical fraud detection: a review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, edited by D. L. Schilling, John Wiley & Sons, New York, NY, USA, 1st edition, 1991.

[12] J. G. Proakis, *Digital Communications*, McGraw-Hill Higher Education, New York, NY, USA, 4th edition, 2000.

[13] J. Robertson, E. W. Tallman, and C. H. Whiteman, "Forecasting using relative entropy," FRB of Atlanta Working Paper No. 2002-22, November 2002, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=355460.

[14] H. Nakahara and S.-I. Amari, "Information-geometric measure for neural spikes," *Neural Computation*, vol. 14, no. 10, pp. 2269–2316, 2002.

[15] D. D. Falconer, F. Adachi, and B. Gudmundson, "Time division multiple access methods for wireless personal communications," *IEEE Communications Magazine*, vol. 33, no. 1, pp. 50–56, 1995.

[16] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[17] M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[18] M. Markou and S. Singh, "Novelty detection: a review—part 2: neural network based approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[19] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings of the SIAM International Conference on Data Mining*, pp. 25–36, San Francisco, Calif, USA, May 2003.

[20] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, "A comparative study for outlier detection techniques in data mining," in *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1–6, IEEE, Bangkok, Thailand, June 2006.

[21] S. Basu and M. Meckesheimer, "Automatic outlier detection for time series: an application to sensor data," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 137–154, 2007.

[22] M. J. Desforges, P. J. Jacob, and J. E. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 212, no. 8, pp. 687–703, 1998.

[23] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," in *Proceedings of the International Conference on Pattern Recognition*, vol. 4, pp. 385–388, IEEE, Quebec City, Canada, 2002.

[24] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, pp. 130–143, IEEE, Los Alamitos, Calif, USA, May 2001.

[25] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.

[26] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[27] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience*, vol. 10, no. 1, pp. 47–69, 2001.

[28] H. Haas and S. McLaughlin, Eds., *Next Generation Mobile Access Technologies: Implementing TDD*, Cambridge University Press, Cambridge, UK, 2008.

[29] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.