ENTROPY FILTER FOR ANOMALY DETECTION WITH EDDY CURRENT REMOTE FIELD SENSORS

By
Farid Sheikhi
May 2014

A Thesis
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of

Master of Applied Science in Mechanical Engineering

Ottawa-Carleton Institute for Mechanical and Aerospace Engineering
University of Ottawa

Master of Applied Science (Submitted in September 2013 - Defended in April 2014)

(The Ottawa-Carleton Institute for
Mechanical and Aerospace Engineering) Ottawa, Ontario

Title: Entropy Filter for Anomaly Detection with Eddy Current Remote Field Sensors
Author: Farid Sheikhi
Supervisor: Dr. Davide Spinello
Co-supervisor: Dr. Wail Gueaieb
Number of Pages: 84

Abstract

We consider the problem of extracting a specific feature from a noisy signal generated by a multi-channels Remote Field Eddy Current Sensor. The sensor is installed on a mobile robot whose mission is the detection of anomalous regions in metal pipelines. Given the presence of noise that characterizes the data series, anomaly signals could be masked by noise and therefore difficult to identify in some instances. In order to enhance signal peaks that potentially identify anomalies we consider an entropy filter built on a-posteriori probability density functions associated with data series. Thresholds based on the Neyman-Pearson criterion for hypothesis testing are derived. The algorithmic tool is applied to the analysis of data from a portion of pipeline with a set of anomalies introduced at predetermined locations. Critical areas identifying anomalies capture the set of damaged locations, demonstrating the effectiveness of the filter in detection with Remote Field Eddy Current Sensor.

Acknowledgements

My first and sincere appreciation goes to Dr. Davide Spinello , my supervisor for his continuous help and support in all stages of this thesis . Without his guidance and inspiration this thesis would not have been possible. I also wish to thank my co-supervisor, Dr. Wail Gueaieb for his stimulating collaboration and perceptive comments on this work.

Finally, I would like to thank Precran and NSERC for their financial support allowing to start the collaboration between InvoDane Engineering and University of Ottawa.

Dedication

I dedicate my thesis to my family, all friends and especially in the memory of my brother, Farhan Sheikhi. A special feeling of gratitude to my loving mother, Fariba Raieszadeh whose words of encouragement and push tenacity ring in my ears.

Finally, I wish to express my great love for my wife, Roxana Ghazi, I thank Roxana for her love, caring, and support; her patience when we were away from each other during this work. Without her help and encouragement it simply never would have been.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

High demand for oil and water transportation is playing an important role in the world economy, as they are primary needs. Pipelines are the most economical way to transport large quantities of water and oil through land or undersea. However, disruption of flow leading to shortage of supply, can lead to high economic losses. Hence, inline inspection and repair are vital for maintaining a pipeline network. On the other hand, oil spills can result in sudden and long-term environmental damage. Considering that millions of miles of oil pipeline networks are operating across the globe, and many more are under construction, [2] automating inspection is crucial. Technology in this field is developing rapidly in order to provide efficient and quick inspection [3,4]. There are two major types of pipe inspection: 1)external pipe inspection, which deals with the exterior of the pipeline. This method is not very common, since pipelines are often underground or there are support beams in the way. 2) internal pipe inspection, generally referred to as pigging, is widely used to guarantee safe and reliable fuel delivery. The majority of Pipeline Inspection Gauge (PIG) technology developers, rely on nondestructive testing (NDT) to quickly and economically detect and characterize degradation and damage. There are many challenges that Inspection tool developers must overcome. They must build inspection tools that survive the pipeline environment, meet specifications for long distance pipelines with high operational speeds, deal with pipeline abnormalities such as tight bends and obstruction in the pipe, and keep the measurement sensors in good condition [5]. To

fit the demand for quick and economical surveys of large portions of pipelines, high speed inspection methods such as visual testing, penetrant testing, magnetic particle testing,magnetic flux leakage, Hall-effect testing radiographic testing, ultrasonic testing, eddy current testing, Thermal infrared testing, magnetic Barkhausen effect, and acoustic emission testing [6–10] are used to fulfill the requirements.

Due to profusion of metallic utility pipes, it is more suitable to use magnetic sensors such as magnetic flux leakage [11], magnetic particle and eddy current [12] as they are capable of detecting both internal and external defects. Magnetic PIGs magnetize the pipe as they travel along it. A magnetic field related signal is captured by a group of transducers that is uniformly distributed around the circumference inside the pipe wall. A difference in the transmitted and received magnetic-dependent signals usually indicates the existence of a flaw near that point [13–16].The flaw can be due to corrosion, weld defects, cracks, fatigue, or deformation. With high noise levels, in order to extract anomalies from the collected sensory data, signal processing techniques, like filtering, are required. Sensory data is characterized by low signal to noise ratio. Therefore, relevant signals associated with features to be identified may be masked by noise.

In this thesis, we investigate the problem of extracting feature from data generated by a multi-channel remote field eddy current sensor. The problem is formulated in the framework of probabilistic classification theory, where data points have to be classified as noise or feature. Some authors developed the idea of using Shannon entropy [17] to filter noisy signals in order to distinguish features from background noise [18–20]. We extend such idea to the data series generated by remote field eddy current sensors by considering the Rényi entropy [21] which is a one-parameter generalization of the Shannon entropy. The characterization of sensor noise and anomalies is done by thresholding the Rényi filter within Neyman-Pearson [22] decision making framework.

In chapter 2, we present a brief survey of the latest techniques in inline inspection tools. In the second section, we provide a comprehensive literature review in the field of anomaly detection and discuss different types of detection. The last section is dedicated to the overview of the most common techniques in anomaly detection and the related applications.

In Chapter 3 we discuss the information entropy, which delineate the theoretical framework of this thesis. Two well-known types of optimal criteria for hypothesis testing, that are Bayes decision criteria and Neyman-Pearson criterion, are discussed. Sensory data in terms of related discrete probability density functions are characterized, as the necessary preamble to apply the entropy filter. The entropy filter is presented and different methodologies on how to obtain optimal bin size for a histogram representation that is linked to the entropy filter are presented. Subsequently, we determine thresholds for hypothesis testing within the Neyman-Pearson decision making framework with statistical parameters intrinsically related to noise and to anomalies to be detected.

In Chapter 4, the algorithm is illustrated by its effectiveness of detecting critical regions associated with known anomalies. Next we demonstrate the efficiency of detecting critical regions using different window and Rényi parameters employed in the previous section. The noisy and anomaly distribution parameters and thresholds and tested on different that are characterized by different anomalies.

Chapter 5 is left for summary and conclusion.

Chapter 2

Background and Literature Review

2.1 Brief Overview of Inline Inspection Tools

Several types of pipeline inspection gauges have been proposed, such as acoustic, ultrasonic sensor [8, 23], and infrared sensor. However, they are limited by deficiency of data-processing techniques, computational complexity, and high cost. Ultrasonic is useful in liquid pipelines for detecting corrosion and cracks [24, 25]. On the other hand, sensors such as flux leakage sensors, magnetic particle sensors, and eddy current sensors have been widely utilized in commercial inline inspection tools, getting benefit from the metallic inheritance of pipelines that allows the pipeline inspection gauges to magnetize the pipe. There is no comprehensive method for finding all types of anomalies, as the current inline inspection tools are limited in the variety of anomalies they can detect [4]. In this chapter, we discuss some of the common nondestructive testing techniques that are commonly used to detect anomalies in pipeline, and discuss their limitations and advantages.

2.1.1 Magnetic Barkhausen Effect

At the beginning of the 20^{th} century Barkhausen (1919) discovered a unique property in ferromagnetic materials. Applying a magnetic force on a ferromagnet results in sudden changes in the magnitude and direction of the ferromagnetic domains. This

sudden change during the magnetization process creates current pulses (noise) in the coil which is winded around the ferromagnetic material. The amount of Barkhausen noise is associated with the amount of impurities, which is a good indicator of flaws. This method is not very costly and is easy to implement; however, it is limited to ferromagnetic materials, mainly iron [9].

2.1.2 Magnetic Particle Inspection

In 1918 Hoke accidentally discovered that magnetic particles, i.e. small metallic pieces, could be used to find locations of defects. In this method, a ferromagnetic piece such as iron, nickel or cobalt is magnetized by direct or indirect magnetization. The presence of flaw or discontinuity in surface or subsurface level of the material results in a magnetic flux leaking. Using ferrous iron particles on the specimen, magnetic flux leakage attracts the particles, which build up in the vicinity of surface anomalies. Although it is simple to use and the devices are cost efficient, this method is limited to ferromagnetic materials only. For getting the best indication the magnetic field should be perpendicular to the direction of the flaw [26]. This method is usually referred to as magnetic particle inspection.

2.1.3 Magnetic Flux Leakage

One of the most common derivations of magnetic particle inspection is magnetic flux leakage. It is one of the most common techniques for non-destructive in-line inspection of pipelines [27]. It is based on permanent magnets that are used to magnetize a sample to its maximum magnetic flux capacity. Regions that have different thickness due to corrosion, weld, crack or deformation perturb the magnetic flux resulting in 'leaking' the flux into the air. Magnetic flux leakage has several advantages over the particle method, one is higher accuracy in detecting anomalies over a confined area. Another advantage is when the inner surfaces of long tubes and pipelines are not easily accessible for visual inspection [28].

2.1.4 Ultrasonic Testing

Ultrasonic testing was inspired by the work of Lord Rayleigh in 1870s on sound waves, "The Theory of Sound". In this study, he discussed the properties of sound waves in different materials, that later led to the development of this technique [6]. In this method, a short ultrasonic pulse-wave with high frequencies is applied to test material to conduct evaluations. This method can be used mainly for flaw detection, dimensional measurement (such as thickness of defects). Ultrasonic testing mainly consists of receiver, transducer, and display device. Generally, the are two types of ultrasonic testing. One is based on reflected waves whereas the other is based on transmitted waves. The reflected-wave method has several advantages over other nondestructive testing methods such as the wide range of operation on different materials other than steel and metals like concrete and wood composites, high penetrating power allowing deep detection into the material, high accuracy and sensitivity for surface and subsurface flaw detection, capability of estimating the dimension, orientation, and shape of defects, and only one side surface needs to be accessible. On the other hand, it has some disadvantages: it requires high skills and training compared to other nondestructive testing evaluations, rough, irregular shape and tiny materials are difficult to inspect, materials that produce high signal noise such as cast iron can create problems in the detection of flaws [29].

2.1.5 Eddy Current

One of the most common and oldest nondestructive testing techniques is based on eddy current sensors [6]. The first idea involving eddy current was introduced in Michael Faraday's work on electromagnetic induction in 1831. His discovery revealed that relative motion between a magnetic field and a conductor (the test object) induces a voltage in the conductor inducing an electric current, called eddy current. In other words, the alternating magnetic field in a primary coil device links to a conducting specimen test, inducing a voltage that results in a current flow in the conducting object as shown in Fig. 2.1.

A few years later, Heinrich Lenz showed that the direction of the induced current in

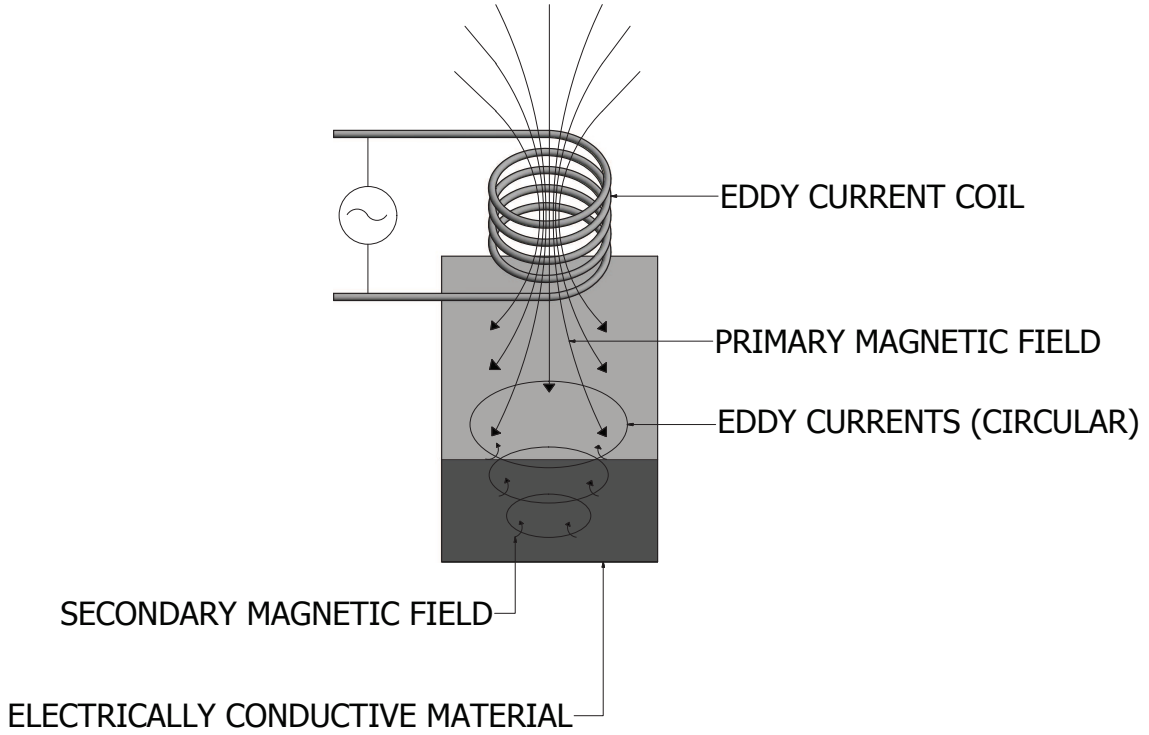


Figure 2.1: Schematic diagram of a simple eddy current testing on a piece of conducting material.

the test object is such that its magnetic field always opposes the magnetic field of the source [6]. The eddy current depends on the change in the inductance of the primary coil (search coil) which is altered in the vicinity of a conducting test object due to an electrical current generated in the specimen when it is subjected to an alternating magnetic field as shown in Fig. 2.2. Hence, by knowing the relation between the magnetic fields of the two coils that generate the eddy current in the conductive material, information can be collected about the test material. The information gained by the test is the electrical conductivity and the magnetic permeability of the test object, the amount of material interfering the coils magnetic field, and the physical condition of the test object such as flaws, defects, cracks etc. Cracks and other surface conditions modify the eddy currents generated in the conductor and give rise to a local change in the impedance. eddy currents always flow parallel to the plane of the winding of the test coil producing them. Thus, a damage parallel to

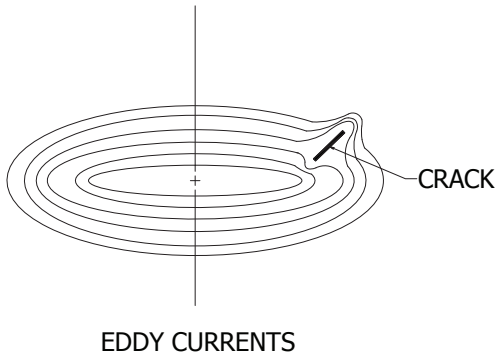


Figure 2.2: Perturbation effect of eddy currents due to defect or discontinuity.

the eddy current can be missed by this method.

2.2 Anomaly Detection

Anomaly refers to patterns in data or signal that do not conform with expected behaviors. A well defined notion of a normal behavior is provided in [1]. Many techniques have been developed in this field for pattern recognition and they are fundamentally similar despite the fact that they appear under different names in the literature, such as outliers, exceptions, surprises, changes, deviation, aberrant, and intrusion. Anomaly detection is crucial because the original data may contain important information. Many studies have shown that anomaly detection is a critical and very challenging task in many safety-required environments such as flow disruption in pipelines, fault detection in critical systems, and fraud detection of credit cards [1, 31–33].

Due to the fact that it is not possible to train a machine learning system to entirely differentiate between normal and anomalous data, it is crucial to investigate the dissimilarities between known and unknown (anomalous) objects. That is why several models for anomaly detection exist which are often tailored to specific types of data. Therefore the choice is dictated by the type of data which is often characterized statistically [32].

The very first attempt for anomaly detection was made by Edgeworth in 1887 as

described in his book "The Method of Measuring Probability and Utility" [34]. Generally speaking, anomaly detection is not the same as noise removal, although they are often related [35]. Noise is considered unwanted data, or not of interest for analysis. It does not belong to any characteristic of an object. On the other hand, anomaly refers to features that could depict some characteristics of an object. Therefore, noise removal does not imply reduction of anomalies. Some authors, such as Aggarwal [36], suggest that anomalies could be considered as a set of noisy data that lies outside of defined regions (clusters) or as a single noise lying outside specified clusters, but in both cases they should be separated from any noise. This means that, generally noise is not considered as an anomaly as its characteristics are available.

2.2.1 Difficulties and Challenges in Anomaly Detection

There are several factors that characterize and pose challenges in the processes that define anomaly detection:

(i) In many cases normal regions are evolving in time, hence there are no fixed boundaries [37].

(ii) It is not an easy task to define a normal region that contains all the possible normal data since the boundary is often not precise and requires a set of purely normal data with which the algorithm can be trained [38].

(iii) The concept of anomaly may vary depending on the application domain. For instance, in medical domains a small deviation of body temperature from the normal temperature might be an anomaly [39]. On the other hand, a similar fluctuation in stock markets might be considered as normal [40].

(iv) It is difficult to obtain labeled data (normal or anomalous) which is accurate and well representative of all types of anomalous and normal behaviors. Naturally, it is more challenging to label anomalous data since it is dynamic and sometimes unpredictable [1, 41, 42].

(v) Noisy (imperfect) data might be confused with anomalies since both are random and do not follow specific patterns [43].

Anomaly detection's success relies on the method as well as the statistical properties of the anomalous data [32]. At the same time, there are several factors that can positively influence anomaly detection [1]:

(a) **Accuracy:** Having low false positive and false negative rates which increase the accuracy for detecting all kinds of anomalous/normal behaviors.

(b) **Simplicity:** Less computationally expensive algorithms require less resources such as limited number of sensors, low memory space and low power transmission. In addition, the number of parameters of the algorithm should be minimized.

(c) **Real-time operation:** It is sometimes essential for any anomaly detection algorithm to be able to work in real-time (online) as in some high level safety critical environments like a spacecraft to avoid hardware/software failures [44].

(d) **Generalization:** The system should be able to generalize without loosing its accuracy and confusing anomaly with normality [45].

(e) **Independence:** A key aspect of anomaly detection is its independence from the attributes of the nature of the input data (also referred to as object, point, sample and entity). Moreover, it should lead to a fairly good performance with a relatively low number of samples in the presence of noise [46]

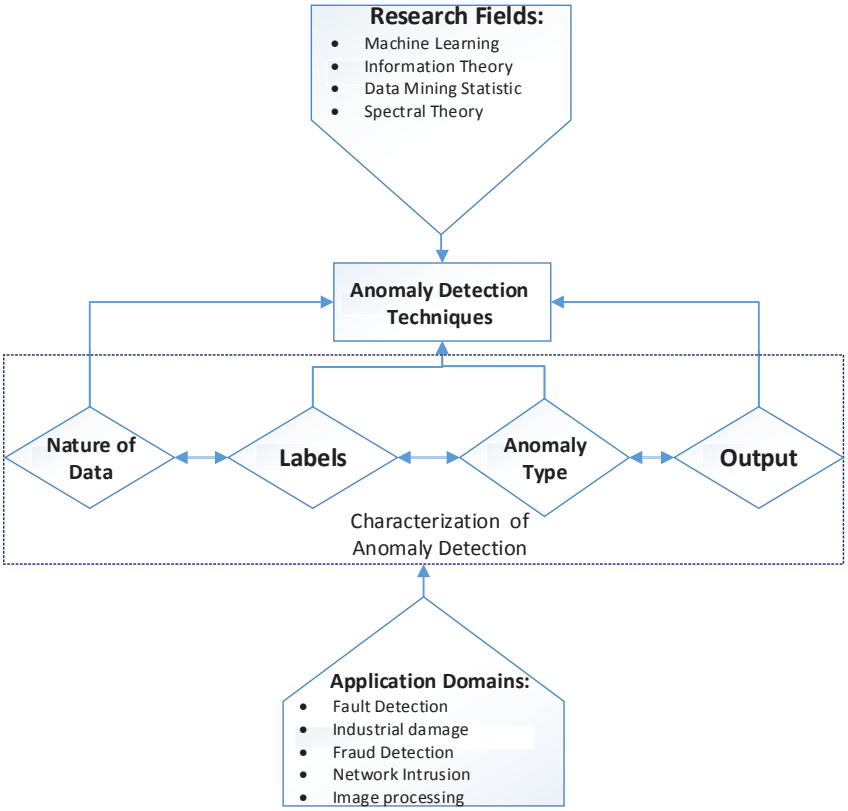


Figure 2.3: Key components of anomaly detection (adopted from [1])

Some of the key components associated with anomaly detection techniques are shown in a block diagram representation, Fig. 2.3. The characterization of anomaly detection techniques is modeled depending on the nature of the input data, availability of the data label, type of anomaly, and output of anomaly detection [1].

2.2.2 Characteristics of Anomaly Detection

Nature of Input Data

Tan *et al.* [46] define input as a collection of data instances also referred to as objects, points, patterns, samples and observations. Each data instance has a feature that can be available in different types like binary, categorical, continuous, and hybrid, where

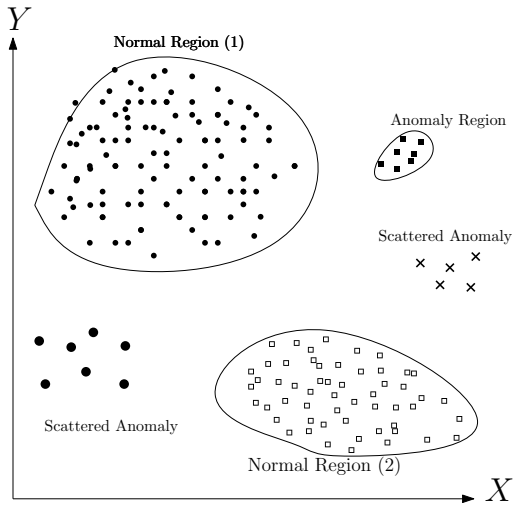
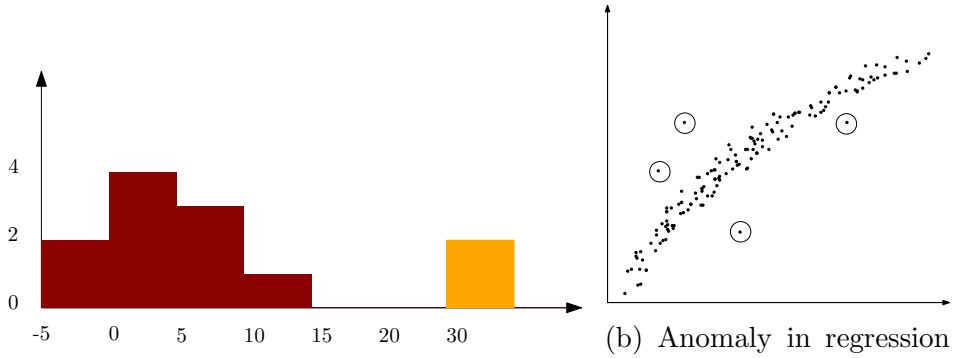


Figure 2.4: Anomalies in multi-class form



(a) Anomaly in histogram form

(b) Anomaly in regression form

Figure 2.5: Different types of anomalies

some data might contain only one feature (univariate) or multiple features (multivariate). One needs to know the nature of data in order to determine the applicability of the anomaly detection method. For instance, in most statistical methods such as nearest neighbor based technique, the nature of attributes would determine the distance measure to be used. In other words, instead of the actual data the relative distance of samples (instances) might be provided. In this case, techniques that require actual instances are not applicable [1].

Types of Anomalies

Another key aspect of this anomaly detection technique is the type of targeted anomaly. Generally, anomalies can be classified into two groups [1].

1. *Point Anomalies* : This is individual data that is anomalous with respect to the rest of the data. This is the simplest type of anomaly and most of the research in this field, including this thesis, is concentrated on this type of anomaly. For instance, in Fig. 2.4, scattered points that lie outside the boundary of the normal regions, are point anomalies since they are distinct from the normal data points. Other types of anomalies are shown in Fig. 2.5 such as anomaly in histogram representation of data and in the regression form. Another example is credit card fraud detection, where a sudden increase in the amount of transactions is considered as an anomaly [47–51].

2. *Contextual Anomalies*: This is individual data that is anomalous within a context. This concept is defined based on the structure in a data set and should be identified in the problem formulation. Each data instance has two main attributes. Contextual attributes, used in spatial data sets where the location longitude and latitude are contextual [52, 53] and in time-series (sequential) where time is contextual and represents the location of an instance in the set [54–56]. The second type is the behavioral attributes, which deal with non-contextual instances, as for example snow in summer (context) could be considered an anomaly (contextual anomaly) but in winter it would be considered normal.

Data Labels

A fundamental part of anomaly detection techniques is labeling the data. Labeling is nothing but denoting data as normal or anomalous. It is computationally expensive to obtain accurate and comprehensive data labels which cover all types of behaviors, especially anomalous data instances due to dynamical nature. There are three types of modes to identify anomalies: supervised, unsupervised, and semi-supervised models.

(a) **supervised:** In this technique, labels are available for both normal and anomalous data. Any unknown data is compared to these two models to decide which class it belongs to. However, there are some issues with this technique. First, usually there is a scarce amount of anomaly data compared to the amount of normal data which results in poor characterization of target events (anomalies) [57–59]. Another issue is that it is impossible to cover all possible ranges of anomaly behaviors since they are always evolving.

(b) **unsupervised:** This method does not assume labels. Hence, it does not require a training data and has a wider application domain. The assumption is based on the fact that anomalies are very rare compared to normal data. Thus, if the assumption is violated it will result in a high false alarm rate [60]. The advantage of this over the supervised method is that it is applicable for on-line anomaly detection as long as incoming data can be classified from a learned model [61].

(c) **semi-supervised:** For this technique labels are available only for normal data (not anomaly). The approach in this method is to build a model that represents the class of normal behaviors from a given training data and then measuring the likelihood of a test instance that is generated by the learned model. The supervised method relies on labeled training data, however training the data in practice may be computationally expensive. On the contrary, the unsupervised method can work without enormous pre-labeling training data, but at the expense of low accuracy. The semi-supervised method provides a trade off between accuracy and adaptivity that reduces the problem to a binary classification [62–65].

Output

There are two possible outputs for anomaly detection techniques: *labels* and *scores*. Labeling assigns an anomalous or normal label to each test instance, while scoring assigns a defect likelihood score to each test instance which indicates the degree of anomalousness. Hence, scoring is more informative than labeling since the analyst can specify a threshold to select the most likely anomalies.

2.3 Techniques and Related Applications to Anomaly Detection

Anomaly detection techniques can be categorized in four main groups *classification based, nearest neighborhood based, clustering*, and *statistical* (parametric and non-parametric) [1]. Fig. 2.6, illustrates the taxonomy of anomaly detection methodologies. For each of the main methodologies, the computational complexity as well as weaknesses and strengths are discussed next.

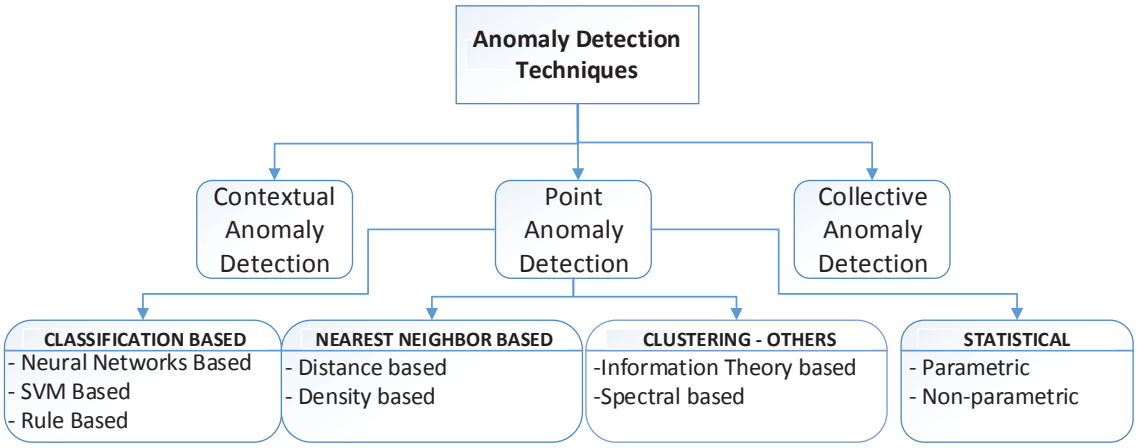


Figure 2.6: Taxonomy of anomaly detection (adopted from [1])

2.3.1 Classification Based

The concept in this technique is to build a classification model for normal and anomalous samples from label data training and then use them to classify each new sample. A sample is considered anomalous if it is not classified as normal by the defined classes. The classification techniques can be divided into two main groups: one-class and multi-class [1]. One-class classification assumes that all training samples have

only one label that is the normal label and attempts to find a separating boundary between the normal data set and the rest of data. The main objective of this method is boundary estimation [66]. Multi-class classification labels data to more than one normal class. One of the widely used techniques is neural networks which is generally a non-parametric approach [31] operating in both single and multiple class settings. There are two stages for a multi-class classification using a neural network. First, it is trained with a normal training data. Then, it is tested by feeding each sample instance to the neural network [67, 68]. Neural classifiers are less affected by noise, and have been extensively applied in machine fault detection applications [69–74], structural damage detection [75–77], as well as image processing [78, 79]. Another classification method is Support Vector Machine (SVM). It constructs a decision hyper-plane as the boundary between classes of given data by maximizing the distance of the closest point to the boundary [80]. A SVM has been implemented in fault detection and machine health prognosis [81–84], power generation plants [72] and image processing [85, 86]. Finally, rule based method which are among the classification techniques which learn rules that detect the normal behavior of a system. If a test instance is not captured by the set rules, then it is treated as an anomaly. In this method each rule has a confidence score that is proportional to the ratio of number of training events correctly classified by the rule to the total number of training events. One of the common types of rule based methods is association rule mining that is being used for one-class anomaly detection by generating rules from the data in an unsupervised environment [1]. This method has been used in fault detection in mechanical units [87], sensor networks, medical applications [88] and public health [89].

Computational Complexity:

The computational complexity of the classification techniques depends on the algorithm that is implemented. Typically, training decision trees tend to be faster while quadratic optimization methods such as SVM are computationally expensive [1]

Strengths:

- The methods have powerful algorithms that can distinguish patterns in multi-class discrimination.
- They can be highly accurate in detecting different kinds of anomalies.

Weaknesses:

- They may be difficult to apply since they require both labels from normal and anomaly classes.
- They can still lead to high false alarm rates in some cases.

2.3.2 Nearest Neighbor Based

The key assumption in this method is that normal data have close neighbors while anomaly data is located far from other data [1, 46]. There are two types of nearest neighbor based methods: distance or similarity based, which specifies anomalies as the data points most distant from the other points using Euclideian or Mahalanobis distance, and density based that identifies anomalies in low density areas. Some applications of this method are pattern recognition [90], medical time series [91], image processing [92, 93] and sensor networks [94, 95].

Computational Complexity:

The computational complexity of this method is in $\mathcal{O}(N^2)$ where N is the data size. Since this kind of technique involves the search of the nearest neighbor for each instance, it does not scale when the number of attributes increases [96].

Strengths:

- It does not require any assumption about the data and is capable of operating with unsupervised and semi-supervised classification paradigms.

- It is fairly easy to adapt this method to a different data types, and it only requires defining measurements (distances) for the given data .

Weaknesses:

- It is computationally expensive.
- If normal or anomalous points do not have a sufficient number of close neighbors, the technique fails to label them correctly resulting in false detections or misses.
- For complex data such as sequences or graphs, it is difficult to select a distance measure that can effectively distinguish between normal and anomalous samples. This is because the performance of this technique highly depends on a distance measure, which is defined between a pair of data instances.

2.3.3 Clustering

Clustering based approaches are based on the unsupervised classification of patterns employed to partition a data set into a number of clusters, where each data instance can be associated to a degree of user-specified level of membership to each of the clusters. The degree of membership can be thresholded to indicate if a data instance belongs (or not) to a cluster. Anomalies can be identified when a instance does not belong to any of known clusters [32]. Clustering based techniques may be similar to the nearest neighbor based techniques in case they require distance computation between a pair of data instances. However, the fundamental difference between the two methods is that clustering based techniques measure each data instance with respect to the cluster that it belongs to, while the nearest neighbor based techniques evaluate each data instance with respect to its local neighborhood [1].

Computational Complexity:

In clustering methods, the computational complexity of processing data depends on the clustering algorithm employed to produce clusters from the data set. Therefore, the computation of such clusters may have $\mathcal{O}(N^2)$ complexity in case the clusters

require pairwise computation of distance for all the data instances [1], or linear using heuristic techniques are used for this purpose [97].

Strengths:

- They can function in an unsupervised mode.
- They can be adopted to other complex data types.
- The evaluation is often fast since the number of clusters are much less than data instances.

Weaknesses:

- The performance of this technique highly depends on the effectiveness of clustering algorithm.
- Some of clustering algorithms might assign every data instance by force to a cluster. This might lead to considering anomalies as normal instances.
- Some of clustering techniques are only effective when anomalies do not form large clusters which could be misinterpreted as normal clusters.

2.3.4 Statistical

Statistical techniques mostly rely on modeling the data based on its statistical properties and then checking if the test samples belong to the same distribution or not [32]. The underlying principle of all statistical approaches is that an anomaly is considered as an observation that is partially or completely irrelevant due to a fact that is not generated by the same assumed stochastic model [98]. The key assumption is that normal data instances occur in high probability regions whereas anomalies occur in low probability regions of the stochastic model [1].
There are few basic approaches that fall under this category. One is based on constructing the density function of known data and then computing the probability of the test data to check if they belong to the normal class. Another simple model

is distance based, where the distance of a data instance that is measured from the estimated mean is the anomaly score for that instance. The threshold is tested on the anomaly score to identify the anomalies [1, 99, 100]. There are different types of distance measures such as Mahalanobis distance, Euclidean distance, Kullback-Leibler distance and Bayesian distance [101, 102]. A more simple statistical model for anomaly detection is box-plot [103] which is a graphical representation of five statistical values (median, lower extreme, lower quartile, upper quartile, upper extreme). Generally speaking, there are two main approaches for probability density estimation, parametric and non-parametric [104]. Parametric methods assume that the data comes from a known underlying distribution [43], whereas non-parametric methods do not generally assume knowledge of an underlying distribution, and build the density function and the parameters from the data [104]. Non-parametric techniques are more applicable since in real world situations there is no or little knowledge of the underlying distribution.

Parametric Techniques

Parametric techniques are often based on the assumption that the data distribution is Gaussian (sometimes Poisson) and can be modeled based on mean and variance (or covariance) [32]. The data distribution can be modeled by building the density function $f(x, \theta)$, where x is the observation and θ is the parameter. One can determine the likelihood of an anomaly by estimating the parameter θ from the observation x (test sample) of a given data. Hence, based on the nature of data (type of distribution) different parametric techniques exist. A Gaussian modeled based method is a common method which assumes that the data is generated from a normal distribution and its parameters are estimated using Maximum Likelihood Estimates. Shewhart [105] declared anomalies to be 3σ distance away from the distribution mean that contains 97.3% of the data samples. Grubb introduced a test for outlier detection using a score $z = \frac{|\bar{x}-x|}{\sigma}$ where σ is the standard deviation, $\bar{x}$ is the mean, and x is the observation. If the value of z is larger than some threshold then the test sample is considered to be an anomaly. Chow [106] investigated the trade-off between the error rate and the rejection rate to determine on optimal threshold to maximize detection with a given

error rate. Chow showed that the best decision rule is to reject a test sample if the maximum posterior-probability is less than a chosen threshold. Hansen *et al.* [107] extended his work by incorporating the role of the classifier confidence to the decision rule. Since in most of real case scenarios the posterior-probability of a data class is not completely known and is affected by errors, Fumera *et al.* [108] resolved this problem by using multiple reject thresholds for each different data class to obtain the optimal decision and reject regions. More improvements have been made in this topic in [67, 109].

Non-parametric Techniques

Non-Parametric Techniques have more flexibility than parametric techniques, since they do not require any assumption on the statistical properties of the data. The simplest and oldest non-parametric method is histogram, which was first introduced by Pearson [110]. Histograms are often used to plot the discrete density function of the data. This method is structured into two main steps. The first step is building the histogram from training data, and the second step is checking if the data samples fall in any of the constructed intervals (bins) . If the sample data does not fall into any bin, the sample is an anomaly. The number of bins represents a trade-off between resolution in sample space and frequency. Choosing a small number of bins creates a flat histogram that cannot capture the shape of the underlying distribution because of the low resolution in the sample space. On the contrary, choosing a large number of bins creates a sharp histogram that suffers from statistical fluctuation, due to the scarcity of samples in each bin. In other words, there is a low frequency resolution [111]. There is no unique optimal bin size as it depends on the actual data distribution and data sample size; however there are various rules for determining the appropriate bin size [111–115].

Computational Complexity:

The computational complexity of statistical anomaly detection techniques depends on the nature of the statistical model. Typically, single parametric distributions

such as Gaussian and Poisson are $\mathcal{O}(N^1)$. Iterative estimation such as Expectation Maximization is linear per iteration but convergence might be slow. Finally, Kernel based estimation is in $\mathcal{O}(N^2)$ in terms of data size N [1].

Strengths:

- It can be utilized to model various types of distributions.
- The anomaly score is associated with confidence interval which is adjustable.
- It can be operated in unsupervised framework without any need of labeled training data [1].

Weaknesses:

- It is difficult to estimate distributions with high dimension.
- Parametric assumptions are very limited in applications.
- Even with reasonable statistical assumption, there are several hypothesis test statistics. Choosing the best one is not simple [116].

2.4 Pipeline Inspection Methods

There are two main types of sensors for pipeline inspection: one is is related to detection inside the pipe and the other is linked to navigation.

The use of sensor for health monitoring of a pipeline is pertinent to the concept of structural health monitoring (SHM). In this process a damage detection and characterization strategy for structures are implemented. There are several papers published on mechanical system diagnostic with emphasis on models and algorithms of data processing targeting specific type of components such as pipelines, cracked shafts, turbomachines, and electrical machines [117–120].

The very first systematic attempt for extraction of features from a damaged structure using measurements (arrays of sensors), and the statistical analysis of these features was proposed by Pau [121]. Pau defined the relation of damage extent and

performance condition in a systematic fashion. He also developed the criteria of condition evaluation based on the reliability analysis together with statistical and probability theories that are practiced in many mechanical structures. Doebling *et al.* issued a comprehensive review of many advanced monitoring technologies that had been further advancing in engineering applications [122]. The paper addressed the importance of the number and location of measurement sensors. It demonstrated several techniques which perform well in most of cases, whereas they work poorly when subjected to measurement constraints. The paper showed that techniques implemented in this field must take into account the limited of number of sensors as well as the take into account physical constraints.

In addition, Sohn *et al.* reviewed different methods related to discriminate features for flaw detection. These methods include statistical discrimination such as Bayesian expression, outlier analysis, neural network, and fuzzy logic [123]. It is important to implement statistical methods to assess whether the changes in features used to identify anomalies (defects) are notable in statistical context. In order to identify the type of defects, data with specific defects must be available from the structure (pipeline) to correlate with the measured features. A supervised learning method is used for this type of studies. The final step, that is the crucial part, is the testing of the available models on actual sensory data to demonstrate the sensitivity of the extracted features to defects and examine the impact of false alarm due to noise [123].

Worden *et al.* studied the problem of damage detection in composite plates using Lamb waves by employing anomaly analysis, neural network, and estimation of probability density function, consecutively. In this method, using outlier analysis the novelty index for each new pattern is defined as the Euclidean distance between the target output and the output from the network data. Subsequently, the probability density function of the feature over the normal condition set is built. This allows the new features in signals to be flagged as anomaly or normality [124].

Chae and Abraham Developed an automated data interpretation system for sewer pipelines. In this system optical data are collected and then multiple neural networks were implemented for feature pattern recognition. Finally, using fuzzy logic the diagnosis of the anomalies in the pipeline is refined. In this paper Fuzzy set theory

techniques were applied in order to automatically identify, classify, and rate pipeline defects by minimizing the error from the neural network system [125].

2.5 Context of the Thesis and Objectives

The main objective of this thesis is to develop, illustrate, and test a data processing and filtering algorithm to detect relevant features that frequently appear in the forms of events with low probability taken from data series with features that may be masked by noise. In order to identify rare events masked by noise, we consider the fact that such events are more ordered, in an information theory context, than the background noise associated to the sensor. Therefore we map the data to the entropy space by means of an entropy filter. The feasibility and effectiveness of the filter using Rényi entropy functions is demonstrated in the context of an anomaly detection environment. Hypothesis validation is carried out following the Neyman-Pearson Lemma.

Chapter 3

The Entropy Filter

One of the tools to approach anomaly detection is using algorithmic entropy in an information theory context. Entropy in information theoretic sense can be interpreted as the average amount of information required to specify the state of a random variable [126]. In this chapter we introduce the theory behind the entropy filter and then we illustrate how the mathematical framework of this entropy filter algorithm is designed and implemented. The entropy filter computes the entropy associated to every data point with a user defined neighborhood of the original data. In order to compute the entropy, a probabilistic measure of the data is derived using histograms to find discrete probability density function. As different number of bins can depict different features of the data, the optimal number of bins for the histograms is determined by an optimization procedure.

3.1 Rényi's Entropy

3.1.1 Shannon Entropy

The very first attempts of uncertainty measurement roots back to Hartley's (1928). He called the number of choices as "amount of measurement" [127]. It was Shannon [128] who came up with the idea of defining the information entropy as a measurement of average uncertainty in a random variable, when the outcome of an information source

is difficult to predict or it is unknown [129]. This provided a more intuitive concept of information and later turned to be the foundation of information theory, and revolutionized the communication fields.

Let X be a random variable with a set of finite events $\{x_1, x_2, x_3, ..., x_N\}$, with cardinality N. Let $P(x_i)$ be a probability mass function over the set X, where $\sum_{i=1}^{N} P(x_i) = 1$. The Shannon entropy is defined as:

$$H_S(X) = -\sum_{i=1}^{N} P(x_i) \ln P(x_i) = -\mathbb{E}[\ln P(X)] \tag{3.1}$$

Consider $H_S(X) \doteq H_N(P_1, ..., P_N)$ where $P_i \doteq P(x_i)$ for $i \in \{1, ..., N\}$ the symbol "$\doteq$" means defined as. Then $H_S(X)$ can be axiomatically characterized as follows [130].

1. *Continuity*: The entropy measure $H_N(P_1, ..., P_N)$ is a continuous function of its arguments.

2. *Permutationally symmetric*: $H_N(p_{i1}, ..., p_{iN}) = H_N(p_1, ..., p_N)$ for every permutation $\{i_1, ..., i_N\}$ of $\{1, ..., N\}$.

3. *Maximality*: The entropy of the uniform discrete distribution is an upper bound for the entropy of any discrete distribution with the same cardinality; $H_N(p_1, ..., p_N) \leq H(1/N, ..., 1/N)$, which is consistent with Jensen inequality [131],

$$H(X) = \mathbb{E}[\ln P(X)] \leq \ln \mathbb{E}[P(X)] = \ln N$$

4. *Recursivity*: The entropy of N outcomes can be expressed in terms of the entropy of $N - 1$ terms, plus a weighted entropy term,

$$H(p_1, ..., p_N) = H(p_1 + p_2, ..., p_N) + (p_1 + p_2) H\left(\frac{p_1}{(p_1 + p_2)}, \frac{p_2}{(p_1 + p_2)}\right)$$

5. *Additivity*: If $p = (p_1, ..., p_N)$ and $q = (q_1, ..., q_N)$ are two independent probability distributions, then the joint entropy can be expressed as, $H(p, q) = H(p) + H(q)$

6. *Expansible*: An event of zero probability does not contribute to the entropy and therefore

$$H_{N+1}(p_1, ..., p_N, 0) = H_N(p_1, ..., p_N)$$

7. *Concavity*: $H(X)$ is a concave function of its argument [132].

8. *Non-negativity*: $H(X) \geq 0$

The entropy of the random variable X is the average amount of information in a message (data). If the event x is highly probable then there is less surprise (or uncertainty) in it's occurrence. Thus, little is learned upon observing x. If x is not likely to occur there is more surprise in occurrence of x and therefore more information is gained by observing. The level of likelihood is measured by the probability distribution $P(X)$ and the quantity $H(X)$ represents the amount of information content. When the data distribution is skewed the entropy value is smaller (because the outcome is more predictable). On the contrary, for symmetric distributions the entropy value becomes higher since there is more challenge to predict the outcome, which results in more randomness in the outcome. To illustrate Shannon entropy, consider the following example: Let X be a random variable of a fair 6-sided die roll (equal probabilities $P(x) = \frac{1}{6}$). The uncertainty of an outcome is

$$H(x) = -\sum_{i=1}^{6} \frac{1}{6} \ln \frac{1}{6} = -\ln \frac{1}{6} = 1.79 \text{ nats (1 bit equals to } \ln 2 \text{ nats)}$$

On the other hand, if the die is not fair with the following probability distribution $\{1/16, 1/8, 1/2, 1/16, 1/6, 1/12\}$ then the entropy is be given by

$$H(x) = -\left[\frac{1}{16} \ln \frac{1}{16} + \frac{1}{8} \ln \frac{1}{8} + \frac{1}{2} \ln \frac{1}{2} + \frac{1}{16} \ln \frac{1}{16} + \frac{1}{6} \ln \frac{1}{6} + \frac{1}{12} \ln \frac{1}{12}\right] = 1.04 \text{ nats}$$

This result indicates that there is less information in a biased die than in a fair die. In other words, the non-uniform distribution has smaller entropy than the uniform distribution.

3.1.2 Rényi Entropy

Alfred Rényi proposed a generalized family of entropies by modifying one of its axioms [133]. Rényi entropy is a parametric family of entropies defined by [21]

$$H_\alpha(X) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{N} P(x_i)^\alpha = \frac{1}{1-\alpha} \ln(\mathbb{E}[P(X)^{\alpha-1}]) \tag{3.2}$$

By using L'Hôpital's rule it can be shown that H_α converges to Shannon entropy at the limit $\lim_{\alpha\to 1} H_\alpha = H_S$

$$\begin{aligned} \lim_{\alpha\to 1} H_\alpha(X) &= \lim_{\alpha\to 1} \frac{\frac{d}{d\alpha} \ln \sum_{i=1}^{N} P(x_i)^\alpha}{\frac{d}{d\alpha}(1-\alpha)} \\ &= \frac{(\sum_{i=1}^{N} \ln P(x_i) P(x_i)^\alpha)(\sum_{i=1}^{N} P(x_i)^\alpha)^{-1} \mid_{\alpha=1}}{-1} \\ &= -\sum_{i=1}^{N} P(x_i) \ln P(x_i) = H_S(X) \end{aligned}$$

Rényi entropy is characterized by the same axioms as Shannon entropy, except that the recursive property does not hold, and the following properties are added [134,135].

1. *Concavity*: For $\alpha \in (0,1)$ Rényi entropy is concave, and for $\alpha > 1$ it is neither concave nor convex.

2. $H_\alpha(X)$ is a bounded and non-increasing function of α.

For the anomaly detection problem studied here, it is preferred to use Rényi entropy with $\alpha < 1$ over Shannon entropy, because for $\alpha < 1$ the smaller probability events contribute more to the entropy value. On the contrary, the data sample with high or low probability do not significantly contribute to the value of entropy. In other words, Rényi information measure is more sensitive to the change in event probability due to its asymmetric distribution. That is to say, the logarithmic average in Shannon information is replaced by average power probabilities [136].

3.2 Partition into Hypothesis Spaces

There are several techniques for determining which models of data generation and measurement are most consistent with a given finite set of data.

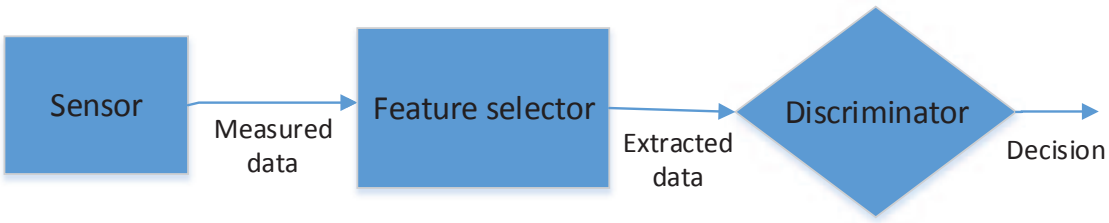


Figure 3.1: Schematic of a basic anomaly detection

Given a random variable H denoting the entropy data set, the goal is to achieve a systematic optimal method to decide which hypothesis generates the data. Consider the problem of identifying a signal masked by noise with the simple binary hypothesis test

$\mathcal{H}_0 : H$ corresponds to the entropy data set associated with sensor noise

$\mathcal{H}_1 : H$ corresponds to the entropy data set associated with known anomalies

Let parameters Θ_0 and Θ_1 be representative of features in the respective hypothesis spaces. We assume that Θ_0 and Θ_1 form a disjoint class of the parameters space Θ. We denote the continuous distributions associated with the two hypotheses as F_0 and F_1 respectively,

$$\mathcal{H}_0 : H \sim F_0 = P(H|\Theta_0) \tag{3.3a}$$

$$\mathcal{H}_1 : H \sim F_1 = P(H|\Theta_1) \tag{3.3b}$$

where $P(H|\Theta_0)$ and $P(H|\Theta_1)$ denote class-conditional probabilities, defined by

$$P(H|\Theta_0) = \int_{\Theta_0} f_0(H)\, dH \tag{3.4a}$$

$$P(H|\Theta_1) = \int_{\Theta_1} f_1(H)\, dH \tag{3.4b}$$

and $f_0(H)$ and $f_1(H)$ are the probability density functions. Let $\pi_0 = P(\mathcal{H}_0) > 0$ and $\pi_1 = P(\mathcal{H}_1) = 1 - \pi_0 > 0$ be a priori class probabilities. Since the sensory data is inherently noisy the parameter chosen to identify anomalies is the entropy associated with data points. Therefore Θ_0 and Θ_1 identify, respectively, the entropy associated with noise and the entropy associated with anomaly.

3.3 Bayes Decision Criteria

A decision rule $\mathfrak{D}$ based on probabilities assigns H to class Θ_1 if

$$P(\Theta_1|H) > P(\Theta_0|H) \tag{3.5}$$

$P(\Theta_i|H)$ denotes a posteriori probability for $i \in \{0, 1\}$ and it may be expressed in terms of the priori probabilities and the class conditional density functions by using Bayes rule

$$P(\Theta_i|H) = \frac{P(H|\Theta_i)P(\Theta_i)}{P(H)} \tag{3.6}$$

Hence the decision rule $\mathfrak{D}$ in (3.5) may be formulated as: select $H \in \mathcal{H}_1$ if

$$P(H|\Theta_1)P(\Theta_1) > P(H|\Theta_0)P(\Theta_0) \tag{3.7}$$

as illustrated in Fig. 3.2

3.4 The Likelihood Ratio Test

If we observe a phenomenon quantified by a random variable distributed according to one of the two distributions in (3.4), we need to decide which of the two best describes it. If $H \in \Theta_i$, then we choose $\mathcal{H}_i$ that is the best fit to the data. Depending on the decision we make (choosing $\mathcal{H}_0$ or $\mathcal{H}_1$) and the true distribution of data, there are four

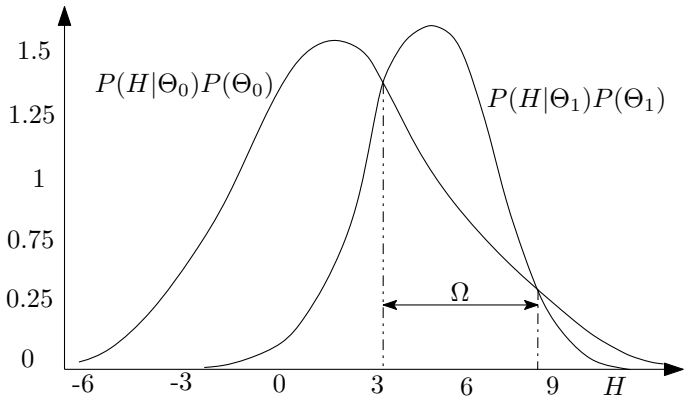


Figure 3.2: Illustration of two Gaussian conditional distributions: if H is in region Ω hypothesis $\mathcal{H}_1$ is selected

	$\mathcal{H}_0$ is true	$\mathcal{H}_1$ is true
$\mathcal{H}_0$ is chosen	(0,0)	(0,1)
$\mathcal{H}_1$ is chosen	(1,0)	(1,1)

Table 3.1: Binary hypothesis test

possible outcomes of the binary hypothesis testing. Let the outcomes be denoted by (i, j) for $i \in \{0, 1\}$ and $j \in \{0, 1\}$, where i represents the decision based on partitions Θ_0 and Θ_1 and j represents the true distribution. The summary of outcomes is shown in Table 3.1.

In order to optimize the decision criteria, we need to introduce a weighting factor for incorrect or correct decisions, generally referred to as loss function, which quantifies how costly each action is. This cost reflects the relative importance of correctly (or incorrectly) deciding on a region (classes). Let $c_{i,j}$ be the cost of choosing $\mathcal{H}_0$ when $\mathcal{H}_1$ (or choosing $\mathcal{H}_1$ when $\mathcal{H}_0$) is true, associated with outcome (i, j) such that the cost of correct decision is less than incorrect decision

$$c_{i,i} < c_{i,j} \quad ; i \neq j \tag{3.8}$$

One can show that the expected loss (Bayes cost) for given decision classes Θ_0 and

Θ_1 is given by [137]

$$\overline{C} = \sum_{i,j=0}^{1} c_{i,j} P_j(\mathcal{H}_j) P(\text{choosing } \mathcal{H}_i; \mathcal{H}_j \text{ is true}) \tag{3.9}$$

where $P_j(\mathcal{H}_j)$ is the priori class probability earlier defined as $\pi_0 = P(\mathcal{H}_0) > 0$ and $\pi_1 = P(\mathcal{H}_1) = 1 - \pi_0 > 0$. In addition, $P(\text{ choosing } \mathcal{H}_i; \mathcal{H}_j \text{ is true})$ refers to the probability of wrong decision, that is choosing $\mathcal{H}_0$ when $H \in \Theta_1$ or choosing $\mathcal{H}_1$ when $H \in \Theta_0$ and can be expressed in terms of the conditional pdf

$$P(\text{ choosing } \mathcal{H}_i; \mathcal{H}_j \text{ is true }) = P(H \in \Theta_i | \mathcal{H}_j \text{ is true }) = \int_{\Theta_i} f_j(H)\, dH \;, i \neq j \tag{3.10}$$

and substituting the above expression into (3.9) yields

$$\overline{C} = \sum_{i,j=0}^{1} c_{i,j} \pi_j \int_{\Theta_i} f_j(H)\, dH \tag{3.11}$$

$$= \int_{\Theta_0} \Big(c_{0,0}\pi_0 f_0(H) + c_{0,1}\pi_1 f_1(H) \Big) dH + \int_{\Theta_1} \Big(c_{1,0}\pi_0 f_0(H) + c_{1,1}\pi_1 f_1(H) \Big) dH \tag{3.12}$$

The integrands are non-negative, thus we need to select regions Θ_0 and Θ_1 such that the cost is $\overline{C}$ minimized. Hence, one should select $H \in \Theta_0$ if the cost being in region Θ_0 is less than of being in Θ_1, that is [137]

$$H \in \Theta_0 \;\; \text{if:} \;\; c_{0,0}\pi_0 f_0(H) + c_{0,1}\pi_1 f_1(H) < c_{1,0}\pi_0 f_0(H) + c_{1,1}\pi_1 f_1(H) \tag{3.13}$$

Similarly if the cost of the second integrand is smaller than the first one,

$$H \in \Theta_1 \;\; \text{if:} \;\; c_{1,0}\pi_0 f_0(H) + c_{1,1}\pi_1 f_1(H) < c_{0,0}\pi_0 f_0(H) + c_{0,1}\pi_1 f_1(H) \tag{3.14}$$

By combining and rearranging equations (3.13) and (3.14) one obtains the simple optimal test:

$$\frac{f_1(H)}{f_0(H)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0(c_{1,0} - c_{0,0})}{\pi_1(c_{0,1} - c_{1,1})} \tag{3.15}$$

For simplicity we assume symmetrical costs

$$c_{1,1} = c_{0,0} \quad \text{and} \quad c_{1,0} = c_{0,1} \tag{3.16}$$

This leads to the likelihood ratio test, in the form

$$\frac{f_1(H)}{f_0(H)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1} \tag{3.17}$$

It can be noted that the term on the right hand side of equation (3.17) depends on prior probabilities, and it is therefore constant with respect to H. This ratio is often denoted as threshold value

$$\eta \equiv \frac{\pi_0}{\pi_1} \tag{3.18}$$

and the left hand side term is the likelihood of H under model f_1 and f_0, denoted as

$$\Lambda(H) \equiv \frac{f_1(H)}{f_0(H)} \tag{3.19}$$

Fig. 3.3 is a graphical representation of the likelihood ratio test for an arbitrary Gaussian distribution.

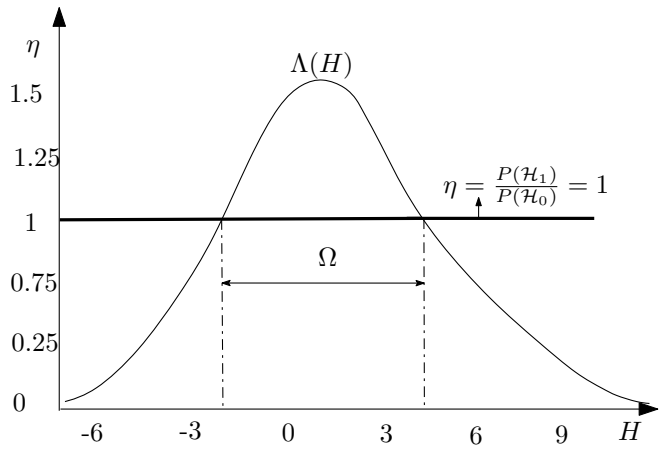


Figure 3.3: Illustration of likelihood function with equal priori probabilities: if $\Lambda(H) > \eta$, the hypothesis $\mathcal{H}_1$ is selected

By rewriting expressions (3.18) and (3.19) in logarithmic form one obtains the log-likelihood ratio test:

	$\mathcal{H}_0$ is true	$\mathcal{H}_1$ is true
$\mathcal{H}_0$ is chosen	N/A[1]	Type II (false rejection)
$\mathcal{H}_1$ is chosen	Type I (false detection)	Detection

Table 3.2: Error terminologies in Neyman-Pearson context

$$\ln \Lambda(H) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta \tag{3.20}$$

3.5 Neyman-Pearson Criterion

The Neyman-Pearson [138] is a binary hypothesis test that does not require prior knowledge of hypothesis probabilities π_0 and π_1. Therefore, Neyman-Pearson is more appropriate because in some applications where it is not reasonable to assign a priori probability to a hypothesis. For instance, we do not know what is the chance of getting hit by a car.

In this context, similar to Bayes decision framework, there are two possible mistakes that may be made in decision process. The first type is declaring $\mathcal{H}_1$ while $\mathcal{H}_0$ is true, and the second type of error is declaring $\mathcal{H}_0$ while $\mathcal{H}_1$ is true. The former misclassifcation is called false detection (false positive) while the latter is called false rejection (false negative), as summarized in Table 3.2.

The average probability of an error is

$$P_e = \beta P(\mathcal{H}_0) + \gamma P(\mathcal{H}_1) \tag{3.21}$$

where β and γ are Type I and Type II errors, defined by

$$\beta = \int_{\Theta_1} f_0(H)\, dH \tag{3.22a}$$

$$\gamma = \int_{\Theta_0} f_1(H)\, dH \tag{3.22b}$$

[1]Historically has been left nameless

It follows that the probability of detection is

$$P_D = 1 - \gamma = 1 - \int_{\Theta_0} f_1(H)\, dH = \int_{\Theta_1} f_1(H)\, dH \tag{3.23}$$

If Θ_1 is the decision region for $\mathcal{H}_1$ and the densities $f_i(H)$ are positive, then as Θ_1 shrinks both the probability of detection and of false detection approach zero. On the other hand, as the decision region Θ_1 expands both probabilities tend to one if f_0 and f_1 tend to overlap. This is a classical trade-off in hypothesis testing.
Neyman-Pearson criterion assigns the value of the threshold by setting a constraint on the probability of false detection and it is formulated as a constrained maximization problem on the probability of detection:

$$\underset{\Theta_1}{max}\{P_D\}, \quad \text{such that } P_F \leq \epsilon \tag{3.24}$$

The maximization is over the decision region Θ_1 and it selects the most powerful test subjected to the constraint on the probability of false detection [101].

Consider the test

$$\mathcal{H}_0 : H \sim f_0(H) \tag{3.25a}$$

$$\mathcal{H}_1 : H \sim f_1(H) \tag{3.25b}$$

where $f_0(H)$ and $f_1(H)$ are density functions. We seek to find the decision rule $\mathfrak{D}$ that maximizes

$$F = P_D + \lambda(P_F - \epsilon) \tag{3.26}$$

$$= \int_{\Theta_1} f_1(H)\, dH + \lambda\Big\{ \int_{\Theta_1} f_0(H)\, dH - \epsilon \Big\} \tag{3.27}$$

$$= \int_{\Theta_1} \big\{ f_1(H) + \lambda f_0(H) \big\}\, dH - \lambda\epsilon \tag{3.28}$$

where λ is a Lagrange multiplier and ϵ is the specified false detection rate. To maximize F with respect to Θ_1 one needs to integrate over H, where the integrand is positive and Θ_1 is set the of H where

$$f_1(H) > -\lambda f_0(H) \tag{3.29}$$

which can be rewritten as

$$\frac{f_1(H)}{f_0(H)} > -\lambda \tag{3.30}$$

This leads to the likelihood ratio test and the decision rule $\mathfrak{D}$ may be formulated as

$$\frac{f_1(H)}{f_0(H)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} -\lambda \tag{3.31}$$

It can be noted that we assumed $\Lambda \neq \eta$. We want to include the case where the likelihood ratio equals the threshold value since it may occur in many problems, especially in tests of discrete data. Therefore, we introduce the function ζ [139]

$$\zeta(H) = \begin{cases} 1 & ; \text{if } \Lambda(H) > \eta \\ \kappa & ; \text{if } \Lambda(H) = \eta \\ 0 & ; \text{if } \Lambda(H) < \eta \end{cases} \tag{3.32}$$

that is the most powerful test of size β, η, κ and has a unique solution subjected to $P_F = \beta$. For the extreme case $\beta = 0$, $\eta \to \infty$, we have $\kappa = 0$. Likewise, for $\beta = 1$, $\eta = 0$ and $\kappa = 1$. To determine the values of η and κ provided that $P(\Lambda(H) = \eta) > 0$, P_F is written as

$$P_F = P(\Lambda(H) > \eta) + \kappa P(\Lambda(H) = \eta) \tag{3.33}$$

The threshold η can then be selected as

$$P\big(\Lambda(H) > \eta\big) \leq \beta \leq P\big(\Lambda(H) \geq \eta\big) \tag{3.34}$$

which allows to select κ by

$$\kappa P\big(\Lambda(H) = \eta\big) = \beta - P\big(\Lambda(H) < \eta\big) \tag{3.35}$$

3.6 Determination of the Optimal Bin Size

A histogram is one of the simplest methods for non-parametric density estimation. It is very efficient when the data sample $x \in \mathbb{R}^d$ (for some dimension $d > 0$) is 1-D

or 2-D [101]. The main issue in histogram modeling is to choose the appropriate number of bins. The number of bins controls the trade-off between the resolution in sample space and frequency. A small number of bins creates flat histograms that cannot represent the shape of the underlying distribution because of low resolution in sample space. On the contrary, a large number of bins creates sharp histograms that suffer from statistical fluctuation due to scarcity of samples in each bin; which leads to low resolution in the frequency [111]. There is no optimal number of bins, as different numbers can reveal different features of the data sample. Therefore, based on the actual data distribution and the aims of analysis different number of bins may be suitable.

There are several rules and guidelines to select the number of bins that compromises between the sampling error and the resolution. Sturges' formula [112] is derived from binomial distribution and is used for normally distributed data. He suggested that the number of bins k is determined by

$$k = \lceil \log_2 N + 1 \rceil$$

where N is the cardinality of the sample space. Doane [113] modified Sturges' formula so that it could be applied to non-normal data by suggesting

$$k = 1 + \log_2 N + \log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right)$$

where g_1 is the skewness of the distribution and

$$\sigma_{g_1} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}$$

Scott [114] proposed a method that is optimal for random samples of normally distributed data, in the sense that it minimizes the Mean Integrated Squared Error (MISE) of density estimate. The optimal bin width (bin size) was proposed to be

$$h = \frac{3.5\hat{\sigma}}{N^{1/3}}$$

and the number of bins can be obtained as

$$k = \left\lceil \frac{max() - min()}{h} \right\rceil$$

where $\lceil\rceil$ is the smallest integer not less than a real number, max( ) is the maximum value of the data, min( ) is the minimum value of the data, and $\hat{\sigma}$ is the sample standard deviation.

Choice based on minimization of an estimated L^2 function

The purpose of this method is to minimize the expected L^2 loss between the histogram and the underlying density function, by estimating the optimal number of bins [111]. The assumption is that data points are sampled independently (Poisson process). In most of classical density estimations, the sampling size is fixed; however, under Poisson assumption the total data size, number of events occurring in a fixed interval, is not fixed but rather obeys Poisson distribution. Therefore, it is more adaptable than the histograms constructed on samples with a fixed number [111].

Algorithm:

(1) Divide the data range into k bins of width h. Count the number of events m_i in the i^{th} bin.

(2) Calculate the mean and variance of the number of events as follows :

$$\mu \equiv \frac{1}{N}\sum_{i=1}^{N} m_i \quad , and \quad \sigma \equiv \frac{1}{N}\sum_{i=1}^{N}(m_i - \mu)^2$$

(3) Compute the cost function:

$$C(h) = \frac{2\mu - \sigma^2}{h^2}$$

(4) Repeat steps 1-3 while varying h to minimize the cost function and find

$$h^\star = \arg\min_h \frac{2\mu - \sigma^2}{h^2} \rightarrow k^\star = \left\lceil \frac{max(x) - min(x)}{h^\star} \right\rceil$$

where $h^\star$ and $k^\star$ are the optimal values of the h and k, respectively.

Summary

In this chapter we discussed Shannon entropy which belongs to the parametric family of entropies know as Rényi entropy. Next we presented a partition of data sets for the purpose of classifying data points as being related to sensor noise or to anomalies. The two classes are associated with a binary hypothesis space with hypotheses denoted respectively by $\mathcal{H}_0$ and $\mathcal{H}_1$. Data classification based on hypothesis testing is implemented within the Neyman-Pearson framework, which allows to compute a threshold that maximizes the probability of detection for a constrained probability of false detection. The final stage was to select to select number of bins to construct a histogram which is used in the next chapter to compute the entropy data set. For this purpose, the most common methods of bin size selection such as Sturges, Doan, Scott, and L^2 function were presented with their algorithm. After evaluation among all the methods, the number of bins for the histogram is determined by an optimization algorithm that is based on minimization of suitable L^2 function.

Chapter 4

Results and Discussion

In data series characterized by low signal to noise ratio, there are rare events that reveal specific features and are masked by noise. The entropy filter is based on the fact that such rare events carry more information than the background noise, and therefore the associated features in the data series can be extracted as they are enhanced by the action of the filter on the original data. In the present study, these features indicate presence of defects (anomalies). We process data series generated by a remote field eddy current sensor [6] mounted on a mobile robotic platform which performs non-destructive inspection of gas pipelines. Sensor measurements are correlated to the presence of defects (anomalies) of the pipeline [140]. Multiple detectors distributed along the circumference of the pipe allow for discretized coverage of the surface. Raw sensory data is characterized by two scalar parameters (i.e. magnitude and phase), as explained in Section 4.1. Based on the indication of the industrial partners that provided the data, we extract anomalies by filtering the phase of the raw data. Entropy filtered values associated to each data point are computed by building the histogram distribution from the data falling within a predetermined neighborhood, with optimal number of bins in the sense discussed in the previous Chapter. The histogram distribution gives the mass probabilities appearing in equation (3.2). In order to classify the filtered data through the Neyman-Pearson criterion, see Section 3.5, we build a probability density function f_0 obtained from a data set acquired from a pipeline without anomalies; thus this data set characterizes signal noise. The classification is

completed by building a probability density function f_1 that characterize anomalies, specifically anomalies artificially introduced on a pipeline to resemble relatively deep dents or losses on material. Predictions of the Neyman-Pearson classifiers with different densities will be illustrated by testing the entropy filter on various data sets with known characteristics.

4.1 Implementation of the filter

In this section we illustrate the entropy filter by showing the different phases on the implementation, starting from the computation of the entropy (that is, mapping the raw data to the entropy space), to the thresholding phase that classifies data points into anomaly and noise. We also show the procedure through which the probability density functions for the Neyman-Pearson classifier are built, by specifically referring to the source data series. The block diagram that summarizes the overall procedure is included in Fig. 4.1.

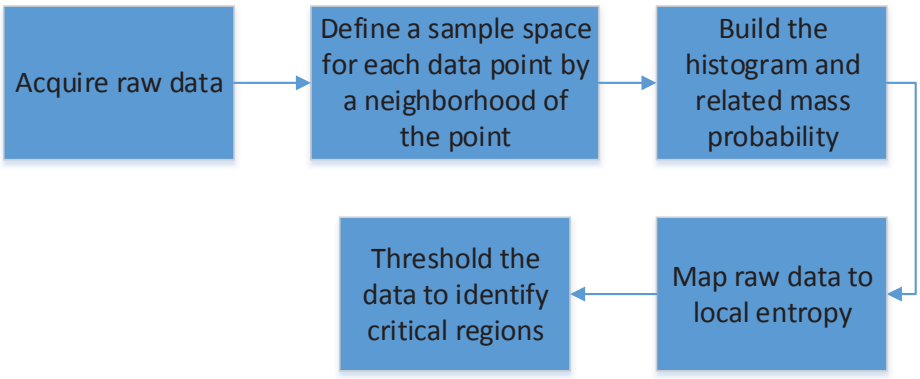


Figure 4.1: Block diagram representation of the overall procedure of the entropy filter

4.1.1 Computation of the Local Entropy

The on-board robotic device operates by taking measurements at a constant velocity. In addition, for this work we consider the torsion of the robot with respect to the axis

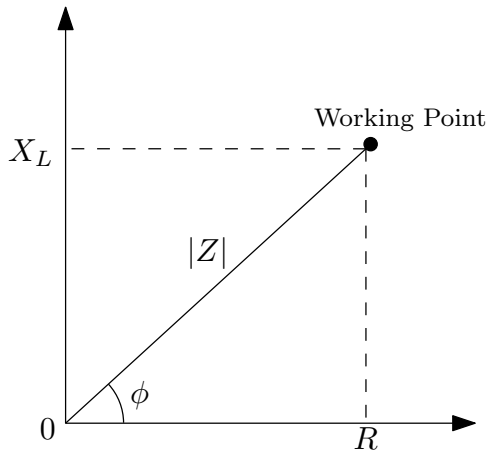


Figure 4.2: Impedence amplitude and phase angle

to be zero as the data analyzed is acquired on portions of pipelines nearly horizontal with no sharp turns. A cross-sectional view of the pipeline with the robot is shown in Fig. 4.3 .

Eddy current sensor functioning depends on the impedance between the electromagnetic field in the sensor coil and the eddy current in the conductive coil. The sensor coil is characterized by the impedance parameter Z which is a complex number defined as

$$Z = R + jX_L \tag{4.1}$$

$$|Z| = \sqrt{R^2 + {X_L}^2} \tag{4.2}$$

where X_L is the inductive resistance and R is the resistive component as shown in Fig.4.2. Raw sensory data from the remote field eddy current sensor is provided in the form of arrays of In-Phase (I) and Quadrature (Q) components, that are basically Cartesian coordinates of the phasor representation of the sensor output. To obtain the phase data we use the relation

$$\phi = \text{atan2}(Q, I) \tag{4.3}$$

where atan2 is the arctangent function with two arguments [141], mapping to $[-\pi, \pi]$ by accounting for the quadrant in which each of the two arguments belong. The scalar field in equation (4.3) is represented as a two dimensional array $\phi(i, j)$, where

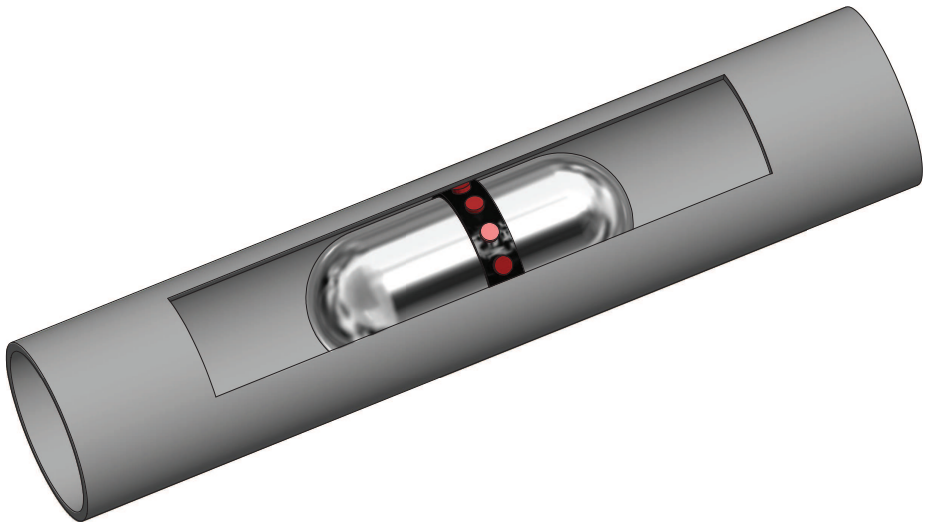


Figure 4.3: A cross-sectional 3D view from portion of pipeline with operating inspection robot (capsule). The rod dots on the robot indicate the eddy current sensors.

i and j span, respectively, the axial position along the pipe and the circumferential position which is sampled by a bundle of channels comprising the sensor, see Fig. 4.6. Phase measurements are correlated to the thickness of the pipe and therefore to the presence of defects (anomalies) quantified by a change in thickness. Since the data is dominated by noise, the aim is to develop an algorithm that enhances the information content so that the anomalies, that are characterized by higher information content than the background noise, can emerge.

The typical output of a channel of the remote field eddy Current sensor is shown in Fig. 4.4. The data series refers to a portion of a pipeline with few significant defects in the middle, and the extremities in which peaks associated with joints appear. Moreover, Fig. 4.5 is provided as a different sample, in contrast with Fig. 4.4 where the signal does not show significant peaks which can clearly be identified. As can be noticed on the exit border there are outliers and these are due to the fact that by exiting the pipe the sensor has taken measurements that are not accurate. For each raw data point $\phi(i,j)$, a subset

$$\Phi(i,j) := \left\{ \phi(h,k) : h \in [i-\ell, i+\ell] \text{ and } k \in \left[j - \frac{w-1}{2}, j + \frac{w+1}{2} \right] \right\} \tag{4.4}$$

that defines a window centered at ϕ_{ij} with sides ℓ and w, see Fig. 4.6. The cardinality

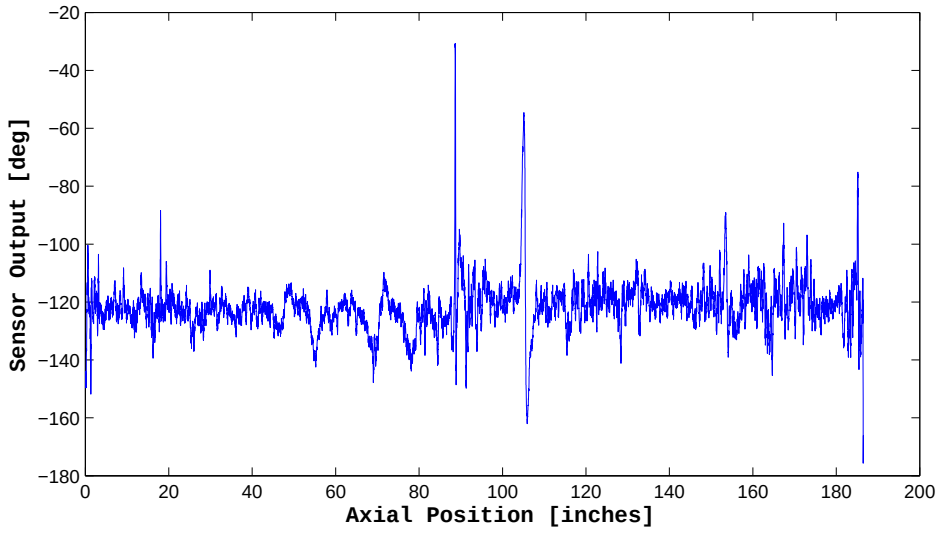


Figure 4.4: Axial position of the output from a channel 2 of the eddy current sensor.

of this data set is therefore $2\ell w$. For this work we choose the size of the local window without having in mind any specific characteristic size of defects, and so we set it in such a way that it is small enough with respect to the overall data size to be able to appreciate local variations, but at the same time, large enough to smooth out small scale oscillations that would not be representative. A systematic study on the correlation between the defect size and window size is left for future work. In order to compute the entropy associated to the data point $\phi(i, j)$, the histogram distribution of the data $\Phi(i, j)$ is built, with sample space given by the set $\Phi(i, j)$. The optimal number of bins is obtained using the minimization of cost function in equation (3.6). To illustrate the process of optimal bin number determination, we select two data points and show in Figs. 4.9 and 4.10 the histograms of with optimal number of bins determined by the minimizers of the plotted cost functions. As expected the optimal number of bins changes as it depends on the cardinality of the sample space as well as the data points. The discrete density function is built by dividing the sample space into N_{s} subintervals $[\phi_k, \phi_k + \Delta\phi_k]_{k=1}^{N_{\mathrm{s}}}$, where $\Delta\phi_k$ is the size of the k^{th} bin and N_s is the optimum number of bins, and by normalizing the frequencies associated to every subinterval with respect to the cardinality and with respect to the size of the

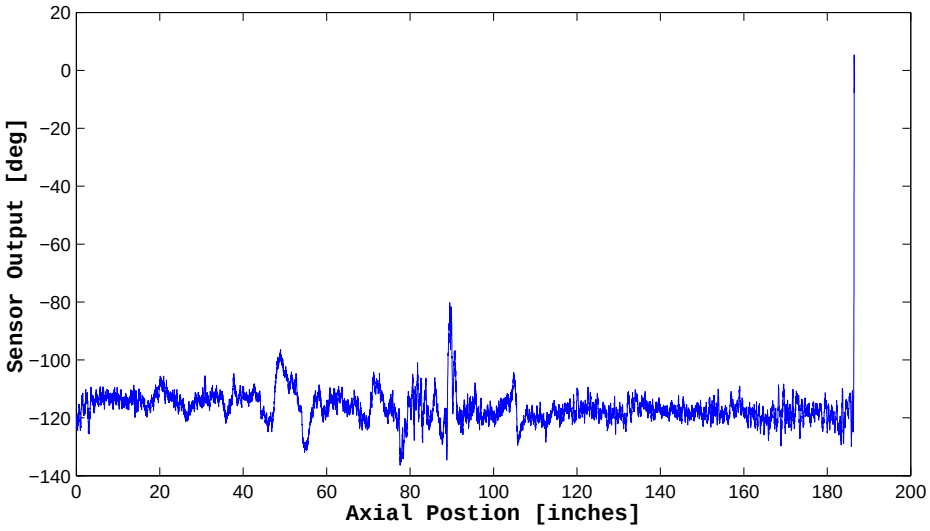


Figure 4.5: Axial position of the output from a channel 5 of the eddy current sensor.

subinterval $\Delta\phi_k$. For the data series in Fig. 4.4 and $N_s = 28$ the discrete probability density function is plotted in Fig. 4.7 and the discrete probability density function of a different data point from the same data series is shown in Fig. 4.8.

Let $p_{ij}(\phi)$ be the discrete probability density function for the data subset Φ_{ij}. For $\phi_{hl} \in [\phi_k, \phi_k + \Delta\phi_k]$ the mass probability $P(\phi = \phi_{hl})$, that is the probability of a sample ϕ taking the value ϕ_{hl} is calculated as

$$P(\phi = \phi_{hl}) = P\left(\phi_{hl} \in [\phi_k, \phi_k + \Delta\phi_k]\right) = p_{ij}(\phi_{hl})\Delta\phi_k \tag{4.5}$$

Using equation (4.5), a discrete probability density function is built, where all the data in a bin has the same probability. For every sample ϕ_{ij} the two dimensional local Rényi entropy is computed by applying the formula

$$H(\phi_{ij}) = \frac{1}{1-\alpha} \ln \sum_{k=1}^{N_s} P(\phi = \phi_k)^\alpha \tag{4.6}$$

In this way a local value of the entropy is associated to each data point. This map from the original data set to the local entropy data defines the entropy filter. The filtered data of Fig. 4.4 is plotted in Fig. 4.12. The output of the entropy filter reflects the fact that the raw data is noisy.

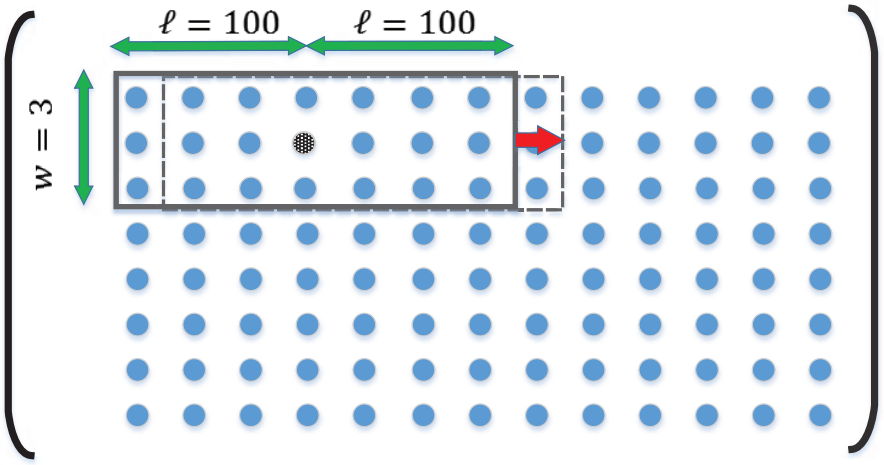


Figure 4.6: Schematic of a rectangular window centered at the phase datum $\phi(i, j)$

4.1.2 Determination of The Threshold

The anomaly detection with the entropy filter is completed with a threshold to discriminate the noise from relevant peaks. The threshold is determined by using the Neyman-Pearson criterion which assigns the threshold value corresponding to the maximum detection probability achievable for a given false detection rate [22, 142]. As evident from equation (3.31) the classification test needs the specification of two probability density functions f_0 and f_1, that characterize respectively the noise and the anomalies, where anomalies could be interpreted as features to be detected.

To determine f_0 we consider a data set from a field trial on a portion of a pipeline that was knowingly anomaly free. The phase data is shown in Fig. 4.11. This data is in cylindrical coordinates and is measured in a lab run from a pipeline with no anomalies. The source of noise comes from non ideal contact of the sensor with the pipe. The figure shows an almost uniform distribution of phase values over the entire space with few light strips that are possibly due to local faulty condition of sensors. The phase data is mapped to local entropy according to the formula (4.6), where each data point is centered on a window defined by parameters $\ell = 100$ and $w = 3$, see Section 4.1.1. The density plot of the entropy data is given in Fig. 4.13. We consider

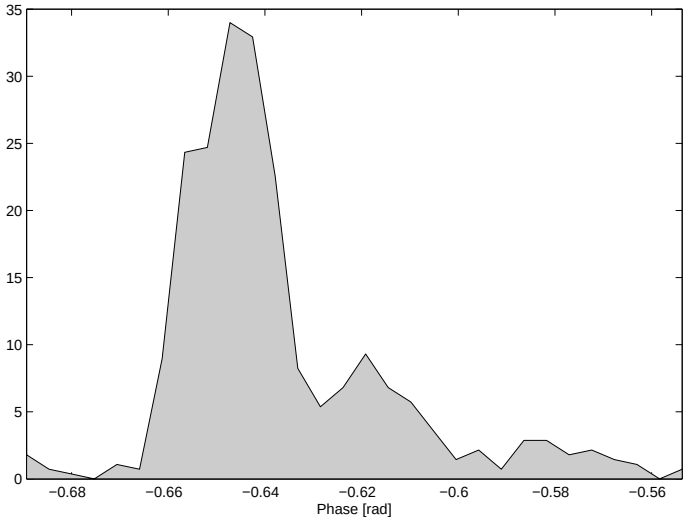


Figure 4.7: Discrete probability density function for the data set in Fig. 4.6 at the indicated data point

the sample space Φ_0 of cardinality N_0, comprised of the values of the computed local entropy, and calculate the first two statistical moments (mean and variance)

$$\mu_0 = \frac{1}{N_0}\sum_{i,j} H(\phi_{ij}^0), \quad \sigma_0^2 = \frac{1}{N_0}\sum_{i,j}\big(H(\phi_{ij}^0) - \mu_0\big)^2 \tag{4.7}$$

where ϕ_{ij}^0 is a data point in Fig. 4.11 belonging to the sample space Φ_0. The plot in Fig. 4.16 shows the discrete probability density function (dots) obtained from the normalized histogram of the data in Fig. 4.13. The continuous line in the same plot is the normal probability density function $\mathcal{N}(\mu_0, \sigma_0)$. In view of the plot in Fig. 4.16, for the sake of simplicity and the fact that the Neyman-Pearson criterion is based on Gaussian distribution we consider the Gaussian density to be an acceptable approximation of the actual discrete density, and therefore assume $f_0 \equiv \mathcal{N}(\mu_0, \sigma_0)$ with $\mu_0 = 3.152$ nats and $\sigma_0 = 0.081$ nats.

The characterization of the density f_1 follows the same steps, except that in this case we consider a data set related to sensor measurements from a portion of pipeline in which there are anomalies that were introduced in a controlled way by machining pipe segments, see Fig. 4.14. Even on a portion of pipeline with artificially

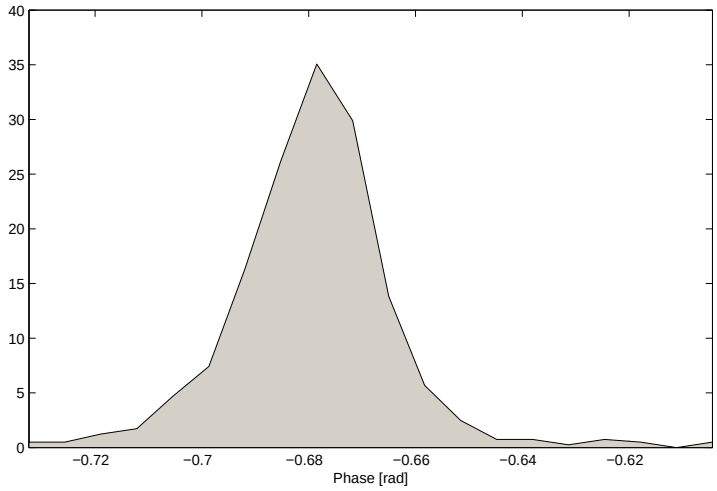


Figure 4.8: Discrete probability density function for the data set in Fig. 4.6 at the adjacent data point.

introduced damages, the number of sensor measurements associated to anomalies is much smaller than the number of sensor measurement associated to undamaged areas (noise). Therefore by including all data points we would obtain statistical moments not significantly different than the ones computed for f_0. Therefore, only the parts that are characteristics of anomalous regions are selected, in order to have a sample space that is representative on anomalies. Specifically, this data set spans eight different anomalies ranging from 0.5 to 3 inch in length and width, and 5% to 30% pipe wall thickness in depth. The density plot of the local entropy computed with $\ell = 100$ and $w = 3$ is given in Fig. 4.15. The discrete distribution from the normalized histogram of the anomaly data in Fig. 4.15 and the approximated Gaussian density are plotted in Fig. 4.17. The parameters of the Gaussian density function are estimated as

$$\mu_1 = \frac{1}{N_1} \sum_{i,j} H(\phi_{ij}^1) \ , \quad \sigma_1^2 = \frac{1}{N_1} \sum_{i,j} \left(H(\phi_{ij}^1) - \mu_1 \right)^2 \tag{4.8}$$

where ϕ_{ij}^1 is a data point in the sample space Φ_1 (with cardinality N_1) comprised of local entropy values extracted to represent anomalies. The computed values of the parameters are $\mu_1 = 2.987$ nats and $\sigma_1 = 0.289$ nats. In this case, since the data

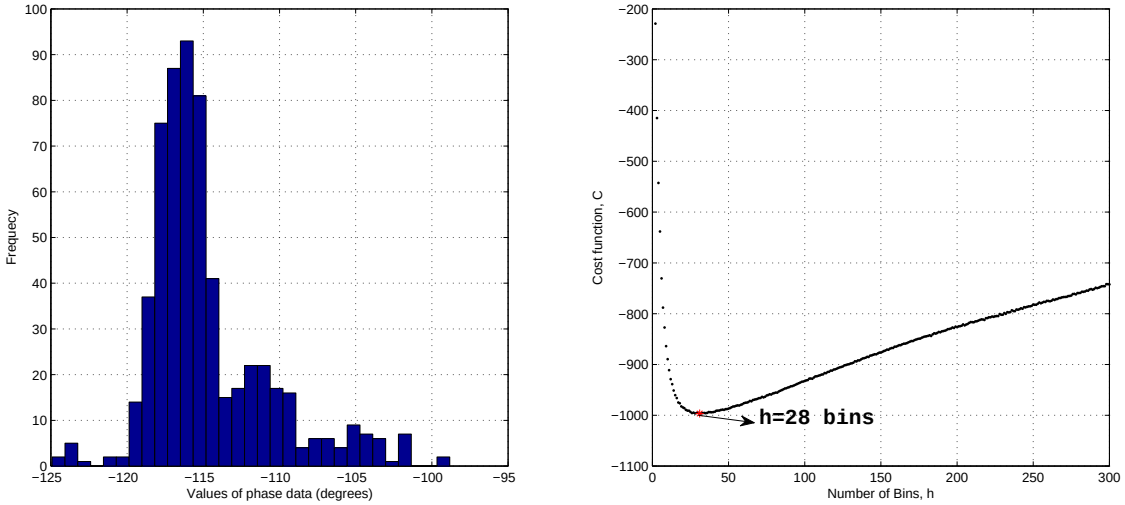


Figure 4.9: Histogram of the sample space of one data point from Fig. 4.4 and related cost function to find the optimal number of bins.

μ_0	σ_0	μ_1	σ_1
3.152	0.081	2.987	0.289

Table 4.1: Parameters for the probability density functions f_0 and f_1 in hypothesis testing.

is not purely noisy, the probability density is not obviously Gaussian; however, for simplicity of treatment we consider f_1 to be approximated by a normal distribution with parameters μ_1 and σ_1, see Fig. 4.17. The statistical parameters in f_0 and f_1 are summarized in Table 4.1. As can be noted the mean of distributions f_0 and f_1 are relatively close making it difficult to distinguish the distributions.

Based on the assumptions for the densities f_0 and f_1, the likelihood ratio of hypothesis testing can be explicitly written as

$$\Lambda(H) = \exp\left(-\frac{1}{2}\left(\frac{H-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{H-\mu_0}{\sigma_0}\right)^2\right) > \eta\frac{\sigma_1}{\sigma_0} \tag{4.9}$$

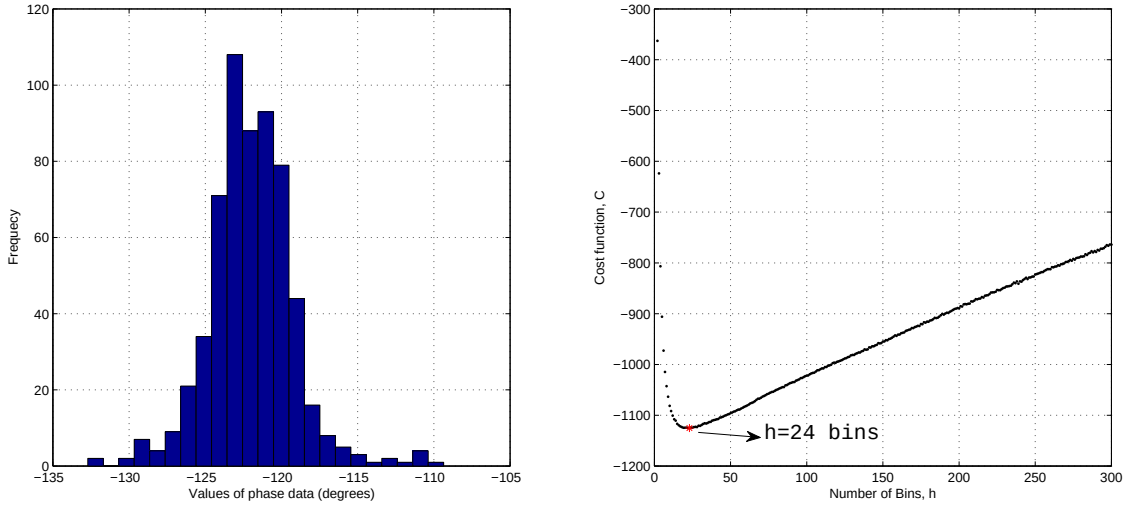


Figure 4.10: Histogram of the sample space of one data point from Fig. 4.4 and related cost function to find the optimal number of bins.

Taking the logarithm of both sides we obtain a quadratic inequality in H that defines the detection region (critical region) as a function of η

$$\left(\frac{H-\mu_0}{\sigma_0}\right)^2 - \left(\frac{H-\mu_1}{\sigma_1}\right)^2 - 2\ln\left(\eta\frac{\sigma_1}{\sigma_0}\right) > 0 \tag{4.10}$$

Let

$$a = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \tag{4.11a}$$

$$b = -2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right) \tag{4.11b}$$

$$c = \frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} - 2\ln\left(\eta\frac{\sigma_1}{\sigma_0}\right) \tag{4.11c}$$

The left-hand side of (4.9) can now be rewritten as $aH^2 + bH + c$. Let $H^-(\eta)$ and $H^+(\eta)$ be the roots of this quadratic function of H, with $H^- \leq H^+$. The detector in

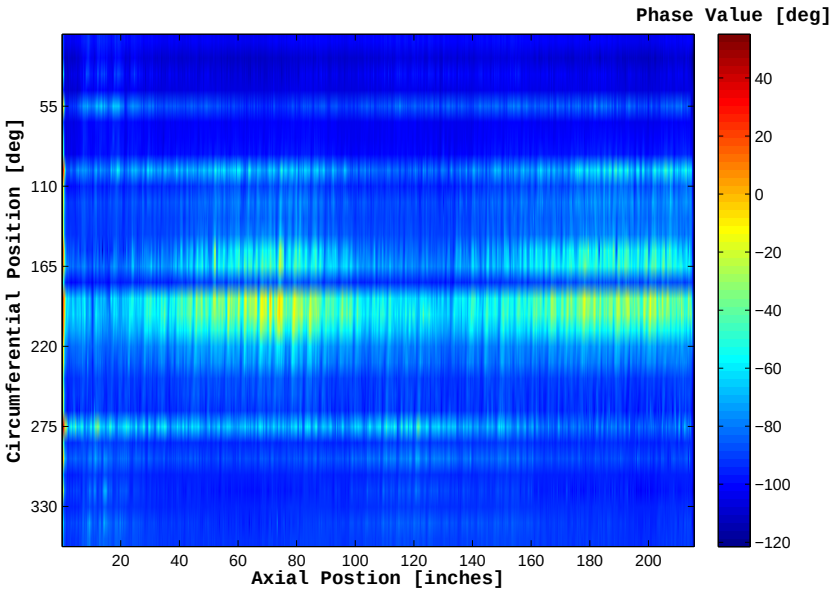


Figure 4.11: 2-dimensional phase data set from noisy signals without anomalies

(4.9) therefore dictates the following detection region

$$\Theta_1(\eta) = \begin{cases} H < H^-(\eta) \cup H > H^+(\eta) & \text{if } a > 0 \\ H^-(\eta) < H < H^+(\eta) & \text{if } a < 0 \end{cases} \tag{4.12}$$

The value of the threshold η is found by numerically solving (3.22a) with the bisection method with assigned $P_F = 5\%$. In applied practice and academic publications, the researcher typically sets the confidence interval at the 95% confidence level which corresponds to 5% of significance level (probability of false alarm), although the choice is largely subjective. Since in this case $a > 0$ as obtained from (4.11a), by using (4.12) equation (3.22a) can be written as

$$G(\eta) := \int_{-\infty}^{H^-(\eta)} f_0(H)\, dH + \int_{H^+(\eta)}^{\infty} f_0(H)\, dH - P_F = 0 \tag{4.13}$$

so that the bisection algorithm finds the root of $G(\eta)$ in the interval $\eta \in [\eta_{min}, \eta_{max}]$, where η_{min} and η_{max} must be provided in such a way that $\text{sign}G(\eta_{min}) = -\text{sign}G(\eta_{max})$ so that there exists $\eta^\star \in \eta \in [\eta_{min}, \eta_{max}]$ such that $G(\eta^\star) = 0$.

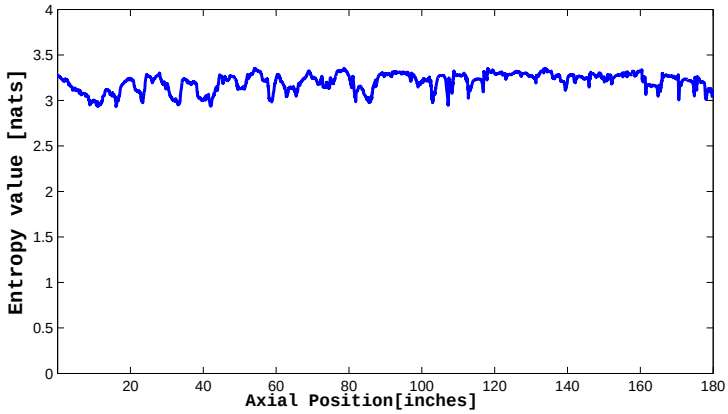


Figure 4.12: Filtered data of a single channel Fig. 4.4 using 2D Rényi entropy explained in section 4.1.1 with $\alpha = 0.5$, $\ell = 100$ and $w = 3$

4.2 Results and Discussions

Filtered data set is considered to belong to a critical region identifying a potential anomaly if $H(\phi_{ij}) \leq H^-$, where the threshold $H^- = 3.004$ is computed as explained in Section 4.1.2, and it depends on f_0 and f_1. The value H^+ is not considered as the condition that anomalies would have higher entropy than noise is not consistent. The entropy filter therefore reduces to the two phases algorithm comprised of the computation of the local entropy and the classification based on the threshold H^-. Comparing Fig. 4.11 and Fig. 4.13 justifies the fact that noisy phase has high information content and therefore is associated to more randomness. The areas with low phase value in Fig. 4.11 (reader might consider them as potential anomalies) are corresponding to high entropy value, see Fig. 4.13, which on the contrary represents noisy signals.

4.2.1 Testing the Consistency of the Algorithm

In order to test the consistency of the entropy filter algorithm, we apply it to the same data sets used to obtain the statistical parameters for the densities f_0 and f_1. What it is expected after thresholding the local entropy data is to obtain almost all

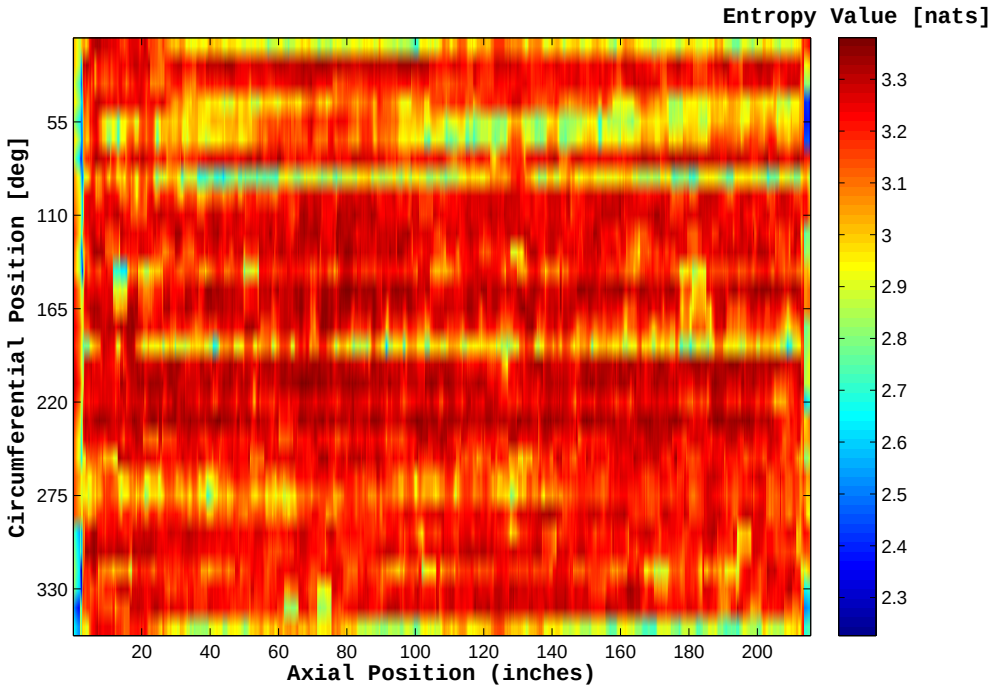


Figure 4.13: Local entropy value data in cylindrical coordinates of Fig. 4.4 using 2D Rényi entropy explained in section 4.1.1 with $\alpha = 0.5$, $\ell = 100$ and $w = 3$

noise in the case of the data set used to compute f_0, and to detect the majority of the anomalies in the data set used to obtain f_1.

The local entropy of the data set used to extract f_0 is computed using $\alpha = 0.5$ and $\ell = 100$ $w = 3$, with results shown in Fig. 4.13. The classification with the threshold H^- obtained through the Neyman-Pearson criterion gives the result in Fig. 4.18, where the output of the entropy filter is represented by the density plot which is obtained by linearly interpolating data from contiguous channels to reconstruct a two dimensional profile, whose support is the surface defining the portion of pipeline. In this case, the 83.4% of the entropy data has value higher than the threshold, and it is therefore classified as noise. Therefore the action of the entropy filter is consistent with respect to the classification of the noise. Results referred to a single channel are shown in Fig. 4.19 to visualize the relation between local entropy data and threshold.

To test the consistency with respect to detection of anomalies from the data set

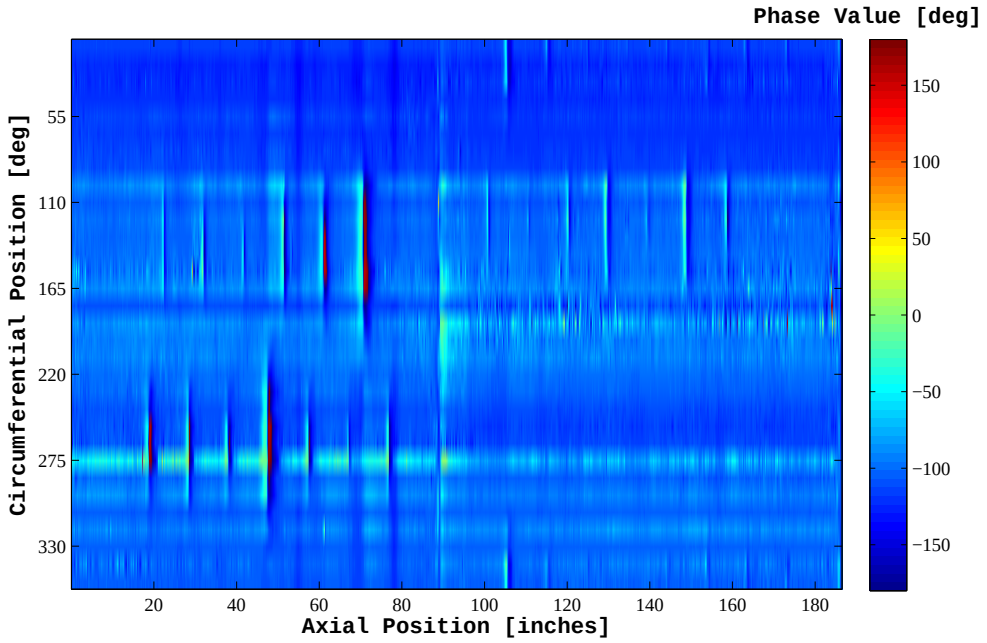


Figure 4.14: Phase data set in cylindrical coordinates from portion of pipeline with known anomalies

used to extract f_1, the entropy filter is applied to a portion of a pipeline across known anomalies. The result is shown in Fig. 4.20, where white areas correspond to critical regions, the entropies of which are below the threshold. The entropy filter clearly captures the anomalies identified by the dots. The output of the filter applied to the all data set gives the density plot in Fig. 4.20. Clearly the critical regions are identified by the algorithm. The relation between the threshold and the entropy data is shown by the plot in Fig. 4.21 that refers to the data series produced by a single channel for the sensor.

The entropy of the data from channel 24 of the sensor in Fig. 4.20 is plotted in Fig. 4.21. The dashed line represents the threshold: values below which, are identified as anomalies.

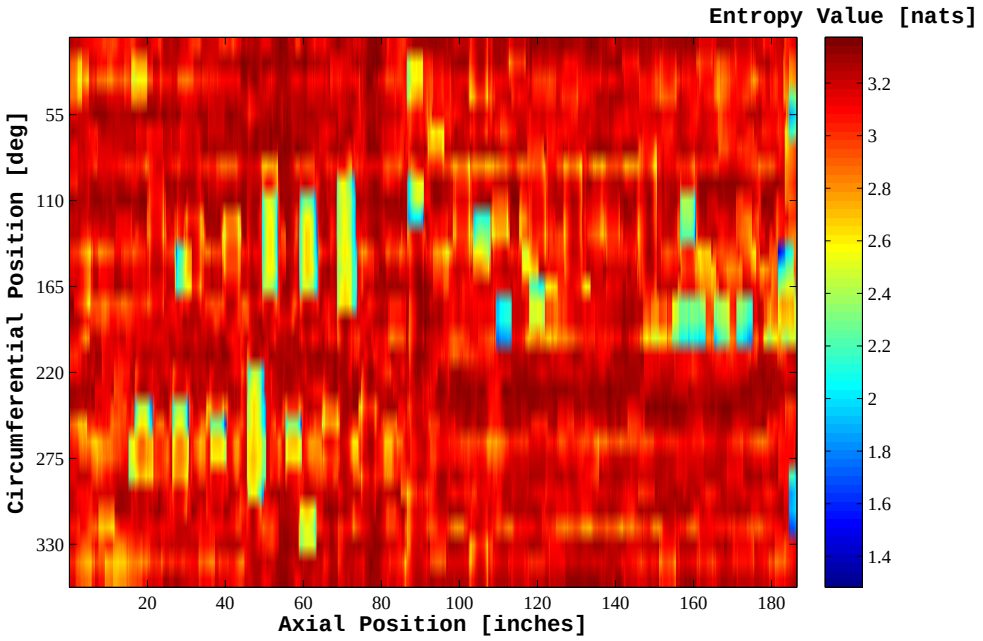


Figure 4.15: Local entropy data of Fig. 4.14 in cylindrical coordinates using 2D Rényi entropy explained in section 4.1.1 with $\alpha = 0.5$, $\ell = 100$ and $w = 3$

4.2.2 Detection of Critical Regions

The entropy filter is applied to data sets not related to the densities f_0 and f_1. Since f_0 and f_1 have been extracted to represent noise and anomalies, it is expected that the entropy filter is capable to classify data series from the same sensor in a fairly accurate way, so that critical regions can be identified and the effort of data analysts can be focused to the detailed analysis of such critical regions.

The application of the entropy filter to a data obtained from measurements on a pipeline with a welded portion is shown in Fig. 4.22. The welded region is clearly identified as anomaly, although in the present form it cannot be distinguished from anomalies of different nature. Ongoing and future work includes the refinement of the algorithm to include multiple features in the classifier, by considering ad hoc densities f_1 extracted from data sets representing specific features to be captured, or by considering a multi-hypotheses testing framework.

Fig. 4.23 shows the output of the entropy filter applied to a data set from a trial on

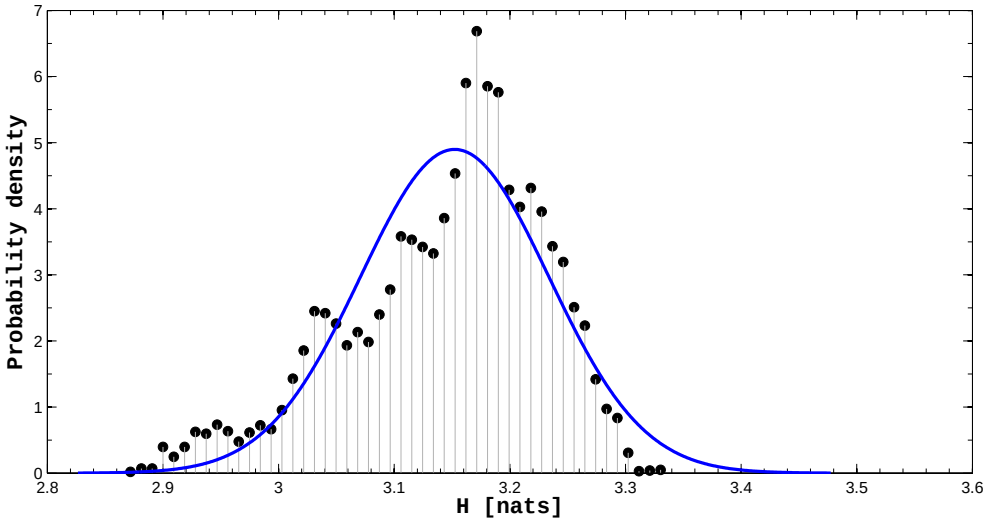


Figure 4.16: Discrete (dots) and Gaussian (continuous line) probability density functions for the entropy data set associated with sensor noise.

a portion of a pipeline with unknown conditions. The filter detects arrays of critical regions along the axis of the pipeline with fairly regular patterns. The regularity of the patterns may be due to the fact that the data is obtained from controlled laboratory tests with artificially generated anomalies; however this information was not provided for this specific data set.

4.3 Qualitative Study of the Influence of Different Parameters on the Rényi Entropy Filter

We study the effect of the Rényi parameter α and the window size parameter ℓ for constant probability of false alarm $P_F = 5\%$ and constant number of channels included in the computation of the local entropy ($w = 3$). The parameter ℓ is varied by taking the discrete values $\{50, 100, 150\}$ while α is varied by taking the discrete values $\{0.1, 0.5, 1.0\}$.

Table 4.2 summarizes the probability of detection (P_D) corresponding to each

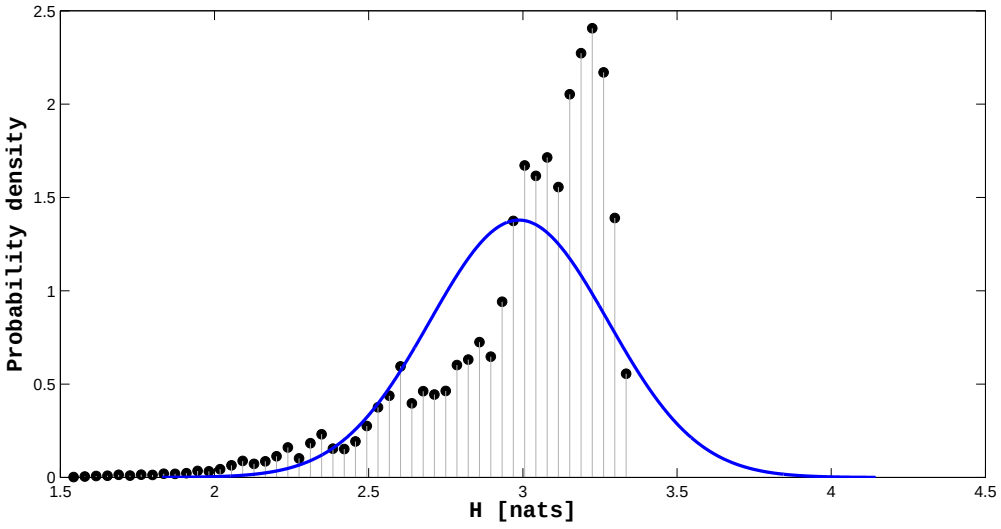


Figure 4.17: Discrete (dots) and Gaussian (continuous line) probability density functions for the entropy data set associated with known anomalies.

	$\ell = 50$	$\ell = 100$	$\ell = 150$
$\alpha = 0.1$	$P_D = 32.8\%$	$P_D = 67.6\%$	$P_D = 83.7\%$
$\alpha = 0.5$	$P_D = 32.5\%$	$P_D = 64.3\%$	$P_D = 79.6\%$
$\alpha = 1.0$	$P_D = 31.4\%$	$P_D = 58.6\%$	$P_D = 74.9\%$

Table 4.2: Probability of Detection using different parameters with fixed false alarm rate $P_F = 5\%$

combination of α and ℓ. The probability of detection increases as the window size increases. This is due to the fact that a larger window size accounts for more data in the neighborhood of a given point, which leads to more evenly distributed probability of phase data, that in turn causes higher entropy value. For fixed statistical parameters that yield constant threshold values, a higher entropy value of each data phase provides higher probability of detection. However, it should be taken into consideration that a larger window size is not necessarily the optimum since it is computationally more expensive.

On the other hand, smaller values of α correspond to a larger probability of

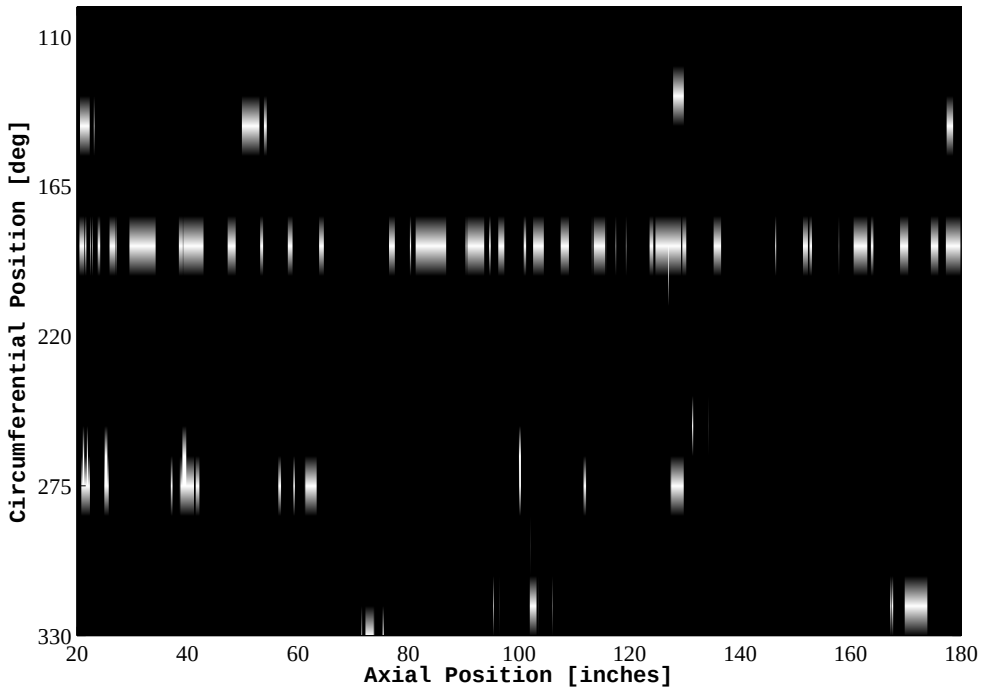


Figure 4.18: Application of the entropy filter to a multichannel data set from a portion of a pipeline with noise and anomaly free. White regions are false critical regions.

detection. As α approaches zero, the Rényi entropy increasingly weighs all possible events more equally, regardless of their probabilities. This is a similar situation to a large window size, as in both cases equal probability of phase data contributes to higher entropy values. As a consequence, with fixed statistical parameters and a fixed threshold value, higher entropy values fall above the threshold, which causes most of the data to be considered as anomaly, which in turn rises the probability of detection. Therefore, besides the computational complexity of larger window sizes, some regions of the pipeline is marked as anomalous, where in fact it is not. This situation rises the false alarm rate and decreases the accuracy of anomaly detection.

The combination of parameters $\alpha = 0.5$ and $\ell = 100$ used in the previous Section gives $P_D = 64.3\%$. Density plots for different combinations of α and ℓ are given in Fig. 4.20 and Figs. 4.24 to 4.27.

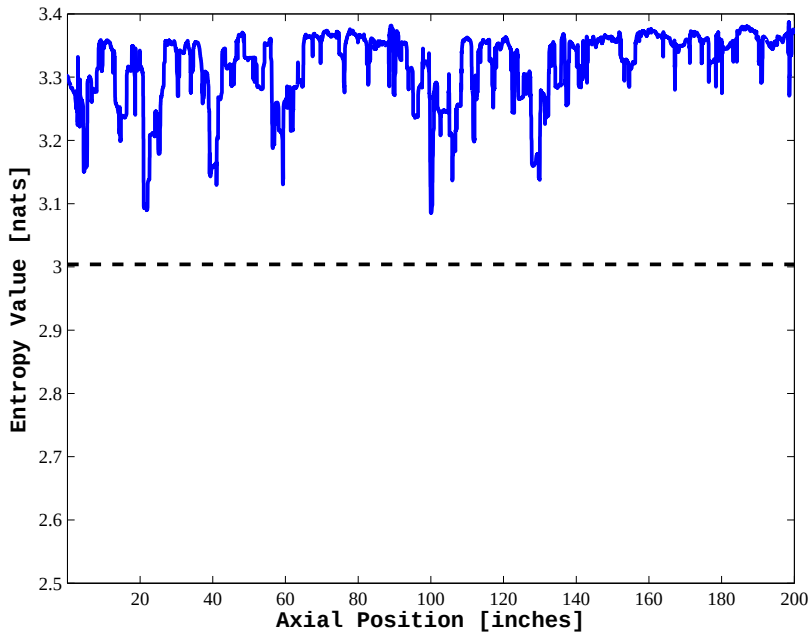


Figure 4.19: Entropy of a single channel output from noisy data set, and the threshold value used with entropy filter.

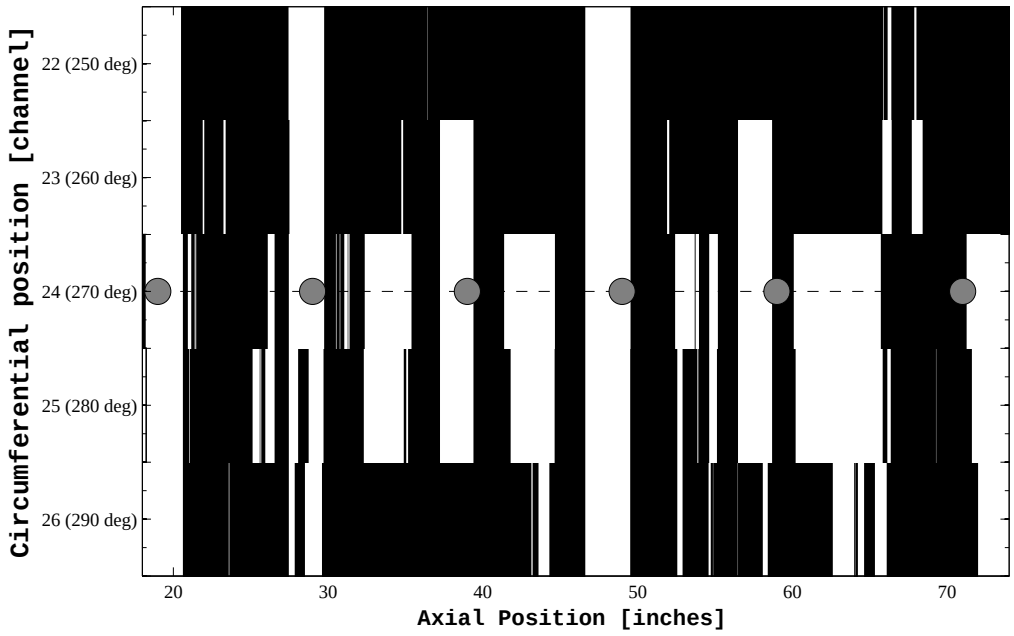


Figure 4.20: Application of the entropy filter to a multichannel data set from a portion of a pipeline with known anomalies (gray circles). White regions are critical with respect to the entropy filter.

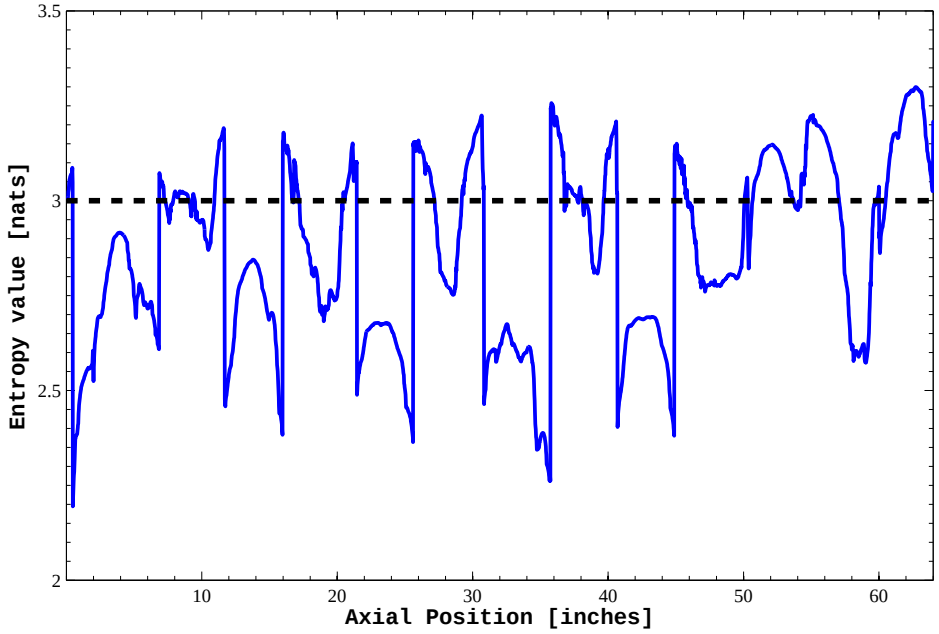


Figure 4.21: Entropy data (continuous line) from channel 24 across a region with known anomalies, and threshold value used with the entropy filter (dashed line).

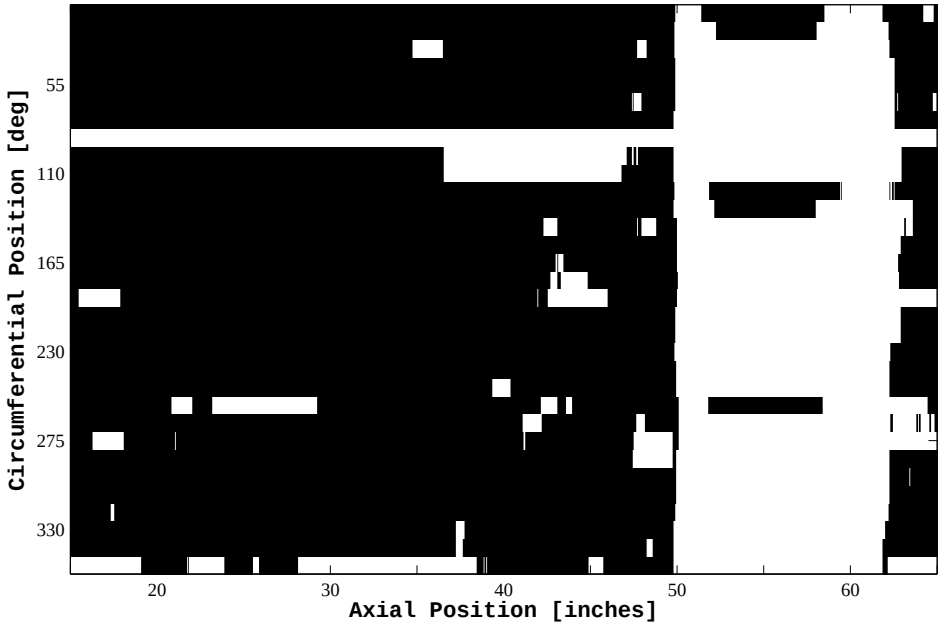


Figure 4.22: Application of the entropy filter to a multichannel data set from pipeline segment with welded joints. White regions are critical with respect to the entropy filter

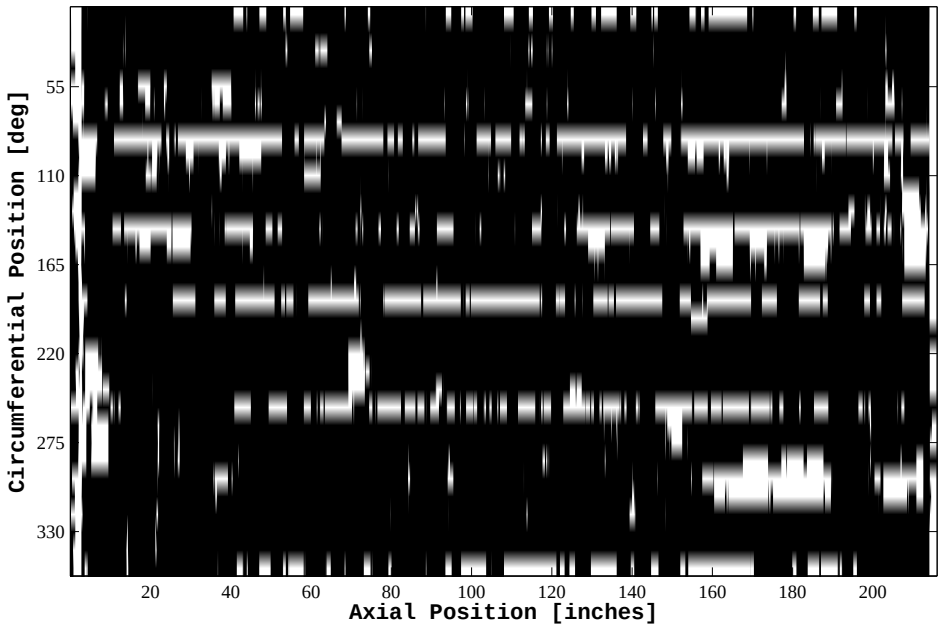


Figure 4.23: Application of the entropy filter to a multichannel data set from a pipeline segment with unknown conditions. White regions are critical with respect to the entropy filter.

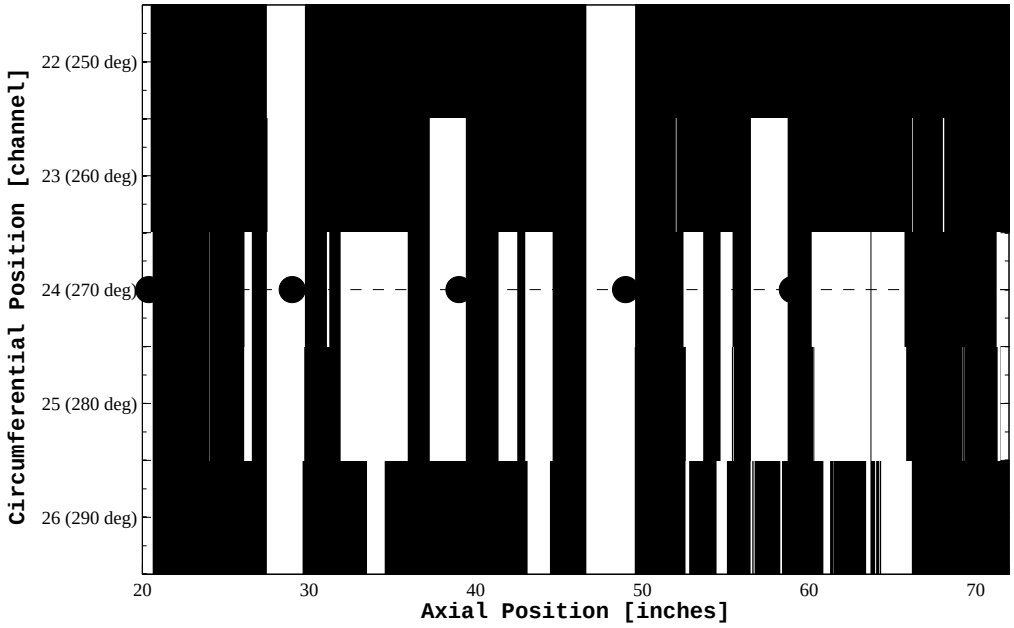


Figure 4.24: Application of the entropy filter with $\alpha = 0.1$ and $\ell = 50$ to a multi-channel data set with thresholds $H^- = 3.015$ and $H^+ = 3.483$

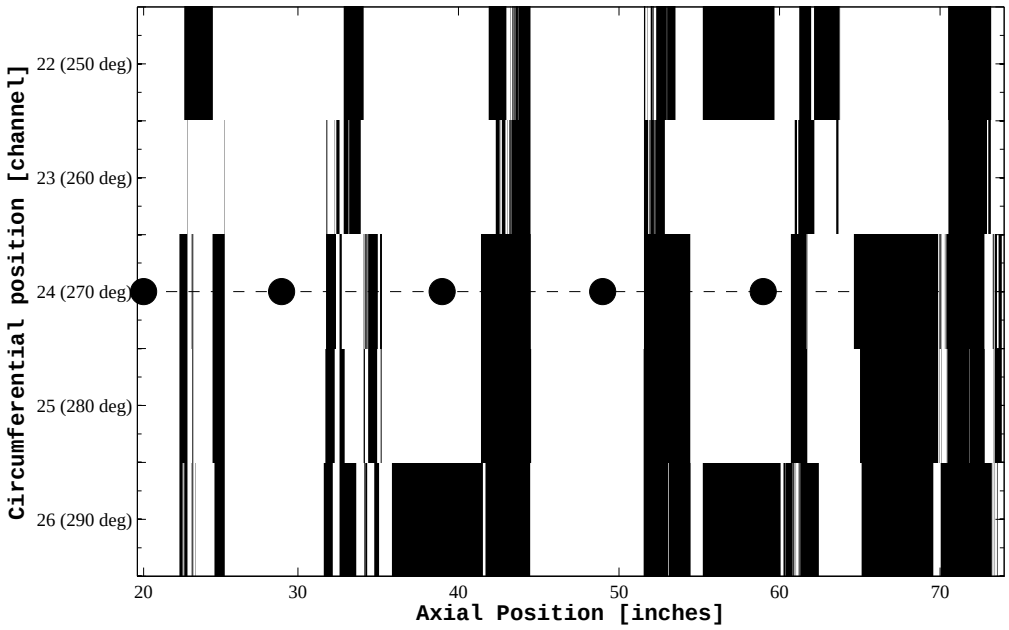


Figure 4.25: Application of the entropy filter with $\alpha = 0.1$ and $\ell = 150$ to a multi-channel data set with thresholds $H^- = 3.846$ and $H^+ = 3.992$

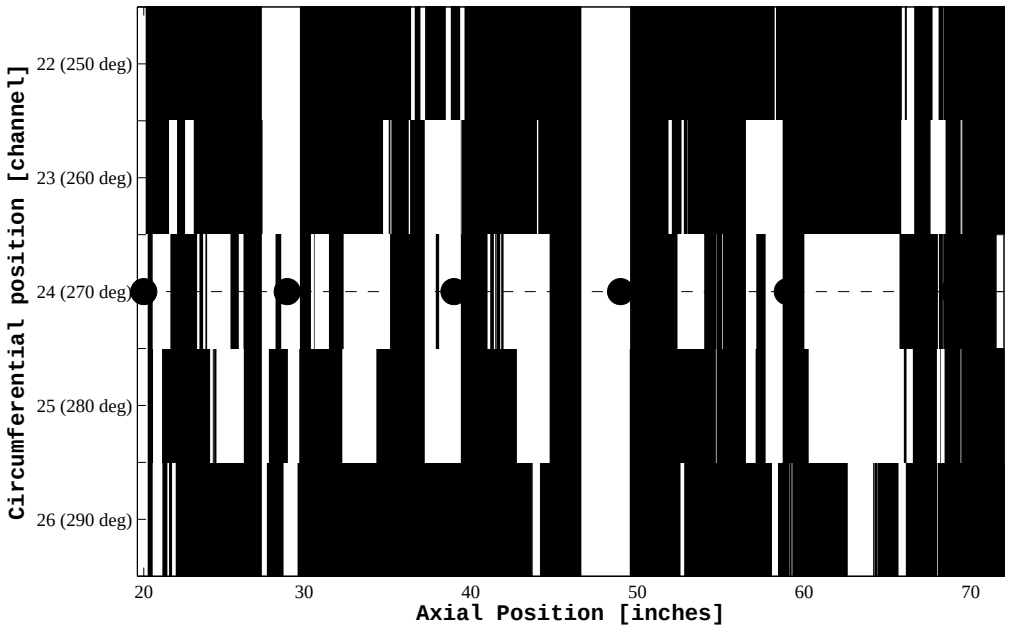


Figure 4.26: Application of the entropy filter with $\alpha = 1$ and $\ell = 50$ to a multichannel data set with thresholds $H^- = 2.592$ and $H^+ = 3.242$

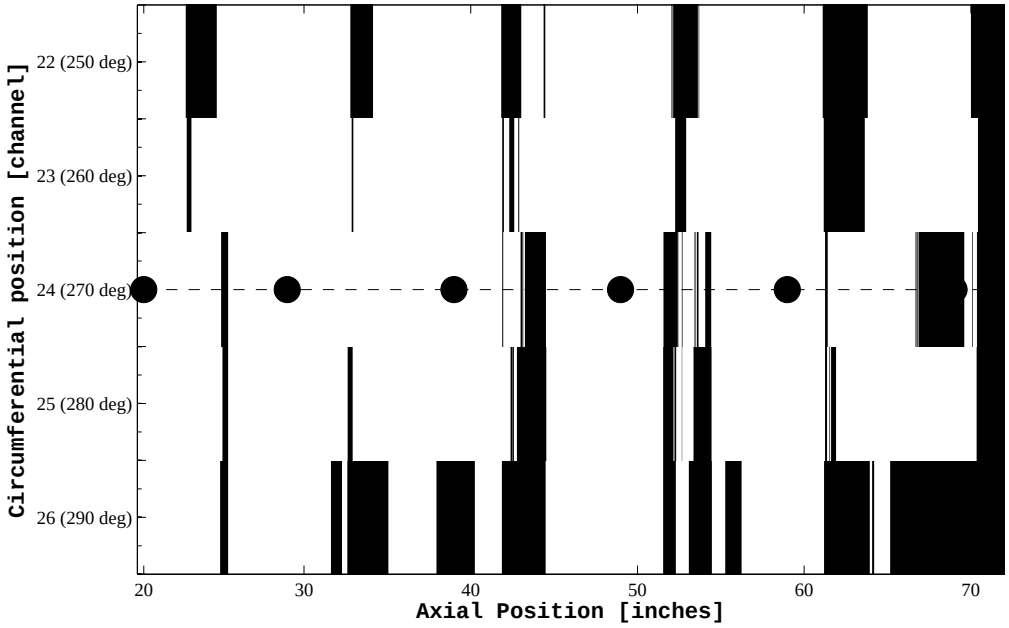


Figure 4.27: Application of the entropy filter with $\alpha = 1$ and $\ell = 150$ to a multichannel data set with thresholds $H^- = 3.375$ and $H^+ = 3.769$

Chapter 5

Summary and Conclusions

This thesis is dedicated to the development, illustration, and testing of a data processing and filtering algorithm to detect relevant features, often in the form of events with low probability, from data series in which such features may be masked by noise. The project originates from the collaboration with a robotic company that develops and maintains a robotic device for nondestructive inspection and damage detection in gas pipelines. The robotic device acquires row data by an on-board remote field eddy current sensor, and the eventual presence of defects can be correlated to the signal characteristics (magnitude and phase) of the data.

In order to identify rare events masked by noise, we take advantage of the fact that such events are more ordered, in an information theoretical sense, than the background noise intrinsically associated to the sensor. Therefore we map the data to the entropy space by means of an entropy filter. The entropy filter computes the entropy associated to every data point, accounting for a user defined neighborhood or subset of the original data. Since the computation of the entropy requires a probabilistic characterization of the data in the neighborhood, a histogram is built to obtain discrete probability density functions that allow to compute the entropy locally. The number of bins for the histograms is determined by an optimization procedure that consists on the minimization of a suitable L^2 norm. Once the original data is mapped into the entropy space, the filter is completed by a classification phase that discriminates between noise and anomalies (features), through a thresholding procedure based on

Neyman-Pearson criterion. Specifically, the threshold that allows to operate the classification is obtained by approximating as Gaussians the distributions of the entropy associated to noise, and of the entropy associated to anomalies. The two distributions are obtained from experimental data sets that respectively are representative of noise and of anomaly, acquired by the robotic device from portions of pipelines with no known anomalies, and with anomalies introduced in a controlled setting.

The operation of the filter is illustrated by testing it on different data sets. As expected, the anomalies introduced in a controlled setting are correctly detected (expected since the entropy representative of anomalies has been obtained from the same data set; therefore this test has been run to check the consistency of the algorithm). In addition, the correlation between phase data and entropy data shows that high phase peaks appear in regions that the filter classifies as anomaly, and it is important test for uniformity of the method. Moreover, the anomaly represented by a weld on a different pipeline is sharply detected, and a pipeline with no known anomalies is predicted to be anomaly free, except the wrong detection associated to one specific channel of the sensor that may be attributed to faulty local conditions of the sensor.

The study of the influence of characteristic parameters of the filter shows the trade-off effect of the neighborhood size on the accuracy of the filter. The observations indicate that too large neighborhoods result into lack of discrimination of data points in terms of information content and into computationally expensive post processing processes. On the other hand, too narrow neighborhoods result in the influence of local fluctuations in the output of the filter. On the other hand, deriving the correlation between size of actual anomalies and the detected ones is beyond the scope of this works and asymmetric study in this direction is left for future work. Overall, the proposed algorithm proves to be effective in introducing a layer of automation in the processing of large sets of experimental data, to help identifying critical regions that eventually may be inspected by experts, or analyzed in a refined way.

Bibliography

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[2] R. Tubb, "2011 worldwide pipeline construction report," *Pipeline and Gas Journal*, vol. 238, no. 1, pp. 18–38, 2011.

[3] L. Udpa, S. Mandayam, S. Udpa, Y. Sun, and W. Lord, "Developments in gas pipeline inspection technology," *Materials Evaluation*, vol. 54, no. 4, 1996.

[4] J. Nestleroth and R. J. Davis, "Application of eddy currents induced by permanent magnets for pipeline inspection," *NDT and E International*, vol. 40, no. 1, pp. 77 – 84, 2007.

[5] J. Nestleroth, "Pipeline in-line inspection challenges to ndt," *insight-wigston then northampton*, vol. 48, no. 9, p. 524, 2006.

[6] C. Hellier, "Handbook of nondestructive evaluation," 2001.

[7] W. Du and S. Yelich, "Post-earthquake pipeline leak detection technologies," *Smart Sensors and Sensing Technology*, pp. 265–283, 2008.

[8] A. Ferrari, "Modelling approaches to acoustic cavitation in transmission pipelines," *International Journal of Heat and Mass Transfer*, vol. 53, no. 19, pp. 4193–4203, 2010.

[9] D. Jiles, "Review of magnetic methods for nondestructive evaluation," *NDT international*, vol. 21, no. 5, pp. 311–319, 1988.

[10] H. Bafl$_a$usen, "Two phenomena revealed with the help of new amplifiers," *Phys Z*, vol. 29, p. 401, 1919.

[11] R. K. Amineh, N. K. Nikolova, J. P. Reilly, and J. R. Hare, "Characterization of surface-breaking cracks using one tangential component of magnetic leakage field measurements," *Magnetics, IEEE Transactions on*, vol. 44, no. 4, pp. 516–524, 2008.

[12] P. Laursen, G. C. Vradis, and C. Swiech, "First robotic device to inspect unpiggable gas transmission pipeline," *Pipeline & Gas Journal*, vol. 236, no. 11, 2009.

[13] R. Ireland and C. Torres, "Finite element modelling of a circumferential magnetiser," *Sensors and Actuators A: Physical*, vol. 129, no. 1, pp. 197–202, 2006.

[14] S. O'Connor, L. Clapham, and P. Wild, "Magnetic flux leakage inspection of tailor-welded blanks," *Measurement Science and Technology*, vol. 13, no. 2, p. 157, 2001.

[15] H. Ramos, O. Postolache, F. C. Alegria, and A. Lopes Ribeiro, "Using the skin effect to estimate cracks depths in mettalic structures," in *Instrumentation and Measurement Technology Conference, 2009. I2MTC'09.* IEEE, 2009, pp. 1361–1366.

[16] T. Theodoulidis, "Analytical model for tilted coils in eddy current nondestructive inspection," *IEEE Transactions on Magnetics*, vol. 41, no. 9, pp. 2447–2454, 2005.

[17] C. E. Shannon, "Prediction and entropy of printed english," *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[18] R. N. McDonough and A. D. Whalen, *Detection of signals in noise.* Academic Press, 1995.

[19] Y. Li, Y. Dong, and G. Sheng, "A signum minimum entropy filter for irregularity detection in optical yarn signals," *Measurement Science and Technology*, vol. 21, no. 3, p. 035104, 2010.

[20] A. Sheinker, N. Salomonski, B. Ginzburg, L. Frumkis, and B.-Z. Kaplan, "Magnetic anomaly detection using entropy filter," *Measurement science and technology*, vol. 19, no. 4, p. 045205, 2008.

[21] A. Renyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[22] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[23] D. Bo, Z. Huiping, S. Sha, and T. Jian, "Research on ultrasonic inspection of pipeline corrosion," in *Control and Automation, 2007. ICCA 2007. IEEE International Conference*. IEEE, 2007, pp. 2572–2575.

[24] K. Reber, M. Beller, H. Willems, and O. Barbian, "A new generation of ultrasonic in-line inspection tools for detecting, sizing and locating metal loss and cracks in transmission pipelines," in *Ultrasonics Symposium*, vol. 1. IEEE, 2002, pp. 665–671.

[25] H. Willems, O. Barbian, and N. Uzelac, "Internal inspection device for detection of longitudinal cracks in oil and gas pipelines- results from an operational experience," in *The 1996 1 st International Pipeline Conference, IPC, Calgary, Can, 06/09-13/96*, 1996, pp. 345–352.

[26] D. Jiles, "Review of magnetic methods for nondestructive evaluation (part 2)," *NDT International*, vol. 23, no. 2, pp. 83–92, 1990.

[27] T. Bubenik, "Magnetic flux leakage technology for natural gas pipeline inspection." Tech. Rep., 1992.

[28] D. Atherton, L. Coathup, D. Jiles, L. Longo, C. Welbourn, and A. Teitsma, "Stress induced magnetization changes of steel pipes–laboratory tests," *on Magnetics IEEE Transactions*, vol. 19, no. 4, pp. 1564–1568, 1983.

[29] D. Orman. (2013, Oct.) Ultrasound-ndt education resource center.

[30] B. Rao, "Challenges and advances in eddy current imaging," *INDE-2007*, 2007.

[31] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

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

[33] I. Lopez and N. Sarigul-Klijn, "A review of uncertainty in flight vehicle structural damage monitoring, diagnosis and control: Challenges and opportunities," *Progress in Aerospace Sciences*, vol. 46, no. 7, pp. 247–273, 2010.

[34] F. Edgeworth, *Metretike: The Method of Measuring Probability and Utility*, 1887.

[35] H. S. Teng, K. Chen, and S. Lu, "Adaptive real-time anomaly detection using inductively generated sequential patterns," in *Research in Security and Privacy, on Computer Society Symposium.* IEEE, 1990, pp. 278–284.

[36] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," *ACM Sigmod Record*, vol. 30, no. 2, pp. 37–46, 2001.

[37] L. Meziou, A. Histace, and F. Precioso, "Alpha-divergence maximization for statistical region-based active contour segmentation with non-parametric pdf estimations," in *Acoustics, Speech and Signal Processing (ICASSP), on International Conference.* IEEE, 2012, pp. 861–864.

[38] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," 2002.

[39] T. SOBERS, "The measurement of internal temperature anomalies in the body using microwave radiometry and anatomical information: Inference methods and error models," Ph.D. dissertation, University of Massachusetts Amherst, 2012.

[40] Z. Fang, G. Luo, S. Xu, and F. Fei, "Stock fluctuations anomaly detection based on wavelet modulus maxima," in *Business Intelligence and Financial Engineering, 2009. BIFE'09. on International Conference*. IEEE, 2009, pp. 360–363.

[41] M. Trochesset and A. Bonner, "Clustering labeled data and cross-validation for classification with few positives in yeast," in *Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics (BioKDD)*, 2004.

[42] J. F. Nieves and Y. C. Jiao, "Data clustering for anomaly detection in network intrusion detection," *Research Alliance in Math and Science*, pp. 1–12, 2009.

[43] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," 2000.

[44] R. R. Lutz and I. Carmen Mikulski, "Operational anomalies as a cause of safety-critical requirements evolution," *Journal of Systems and Software*, vol. 65, no. 2, pp. 155–161, 2003.

[45] D. Tax and R. Duin, "Outlier detection using classifier instability," *Advances in Pattern Recognition*, pp. 593–601, 1998.

[46] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Minig*, ser. Pearson international Edition. Addison Wesley, 2006.

[47] S. P. Singh, S. S. P. Shukla, N. Rakesh, and V. Tyagi, "Problem reduction in online payment system using hybrid model," *arXiv preprint arXiv:1109.0689*, 2011.

[48] R. J. Bolton, D. J. Hand *et al.*, "Unsupervised profiling methods for fraud detection," *Credit Scoring and Credit Control VII*, pp. 235–255, 2001.

[49] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 721–726, 2008.

[50] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *Intelligent Systems and their Applications, IEEE*, vol. 14, no. 6, pp. 67–74, 1999.

[51] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*. IEEE, 1999, pp. 103–106.

[52] P. Chhabra, C. Scott, E. D. Kolaczyk, and M. Crovella, "Distributed spatial anomaly detection," in *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE, 2008, pp. 1705–1713.

[53] J. Mennis and D. Guo, "Spatial data mining and geographic knowledge discovery—an introduction," *Computers, Environment and Urban Systems*, vol. 33, no. 6, pp. 403–408, 2009.

[54] V. Jakkula and D. J. Cook, "Anomaly detection using temporal data mining in a smart home environment," *Methods of information in medicine*, vol. 47, no. 1, pp. 70–75, 2008.

[55] T. Lane and C. E. Brodley, "Temporal sequence learning and data reduction for anomaly detection," *ACM Transactions on Information and System Security (TISSEC)*, vol. 2, no. 3, pp. 295–331, 1999.

[56] N. Ye *et al.*, "A markov chain model of temporal behavior for anomaly detection," in *Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, vol. 166. Oakland: IEEE, 2000, p. 169.

[57] G. M. Weiss and H. Hirsh, "Learning to predict rare events in event sequences," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 359–363.

[58] R. Vilalta and S. Ma, "Predicting rare events in temporal domains," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on.* IEEE, 2002, pp. 474–481.

[59] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[60] D. Guthrie, L. Guthrie, B. Allison, and Y. Wilks, "Unsupervised anomaly detection," in *Proceedings of the twentieth international joint conference on artificial intelligence (IJCAI 2007)*, 2007, pp. 1626–1628.

[61] M. Amer and S. Abdennadher, "Comparison of unsupervised anomaly detection techniques," Ph.D. dissertation, Bachelor's Thesis 2011, http://www. madm. eu/_media/theses/thesis-amer. pdf, 2011.

[62] R. Sillito and R. Fisher, "Semi-supervised learning for anomalous trajectory detection," in *Proc. BMVC*, 2008, pp. 1035–1044.

[63] G. Xiang and W. Min, "Applying semi-supervised cluster algorithm for anomaly detection," in *Information Processing (ISIP), on Third International Symposium.* IEEE, 2010, pp. 43–45.

[64] L. Gao, A. Grant, I. Halder, R. Brower, J. Sevransky, J. P. Maloney, M. Moss, C. Shanholtz, C. R. Yates, G. U. Meduri *et al.*, "Novel polymorphisms in the myosin light chain kinase gene confer risk for acute lung injury," *American journal of respiratory cell and molecular biology*, vol. 34, no. 4, p. 487, 2006.

[65] T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai, "Semi-supervised detection of collective anomalies with an application in high energy particle physics," in *Neural Networks (IJCNN), on The 2012 International Joint Conference.* IEEE, 2012, pp. 1–8.

[66] V. Roth, "Outlier detection with one-class kernel fisher discriminants," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, 2004.

[67] C. De Stefano, C. Sansone, and M. Vento, "To reject or not to reject: that is the question-an answer in case of neural classifiers," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions*, vol. 30, no. 1, pp. 84–94, 2000.

[68] O. Taylor and D. Addison, "Novelty detection using neural network technology," in *COMADEM 2000: 13 th International Congress on Condition Monitoring and Diagnostic Engineering Management*, 2000, pp. 731–743.

[69] I. Loboda and Y. Feldshteyn, "Polynomials and neural networks for gas turbine monitoring: a comparative study," *International Journal of Turbo and Jet Engines*, vol. 28, no. 3, pp. 227–236, 2011.

[70] C. M. Bishop, "Novelty detection and neural network validation," in *Vision, Image and Signal Processing, IEEE Proceedings-*, vol. 141, no. 4. IET, 1994, pp. 217–222.

[71] C. C. K. P. Bennett, "A linear programming approach to novelty detection," *Advances in neural information processing systems*, vol. 13, p. 395, 2000.

[72] S. King, D. King, K. Astley, L. Tarassenko, P. Hayton, and S. Utete, "The use of novelty detection techniques for monitoring high-integrity plant," in *Control Applications, 2002. Proceedings of the 2002 International Conference*, vol. 1. IEEE, 2002, pp. 221–226.

[73] T. Petsche and S. J. Hanson, "Neural network auto-associator and method for induction motor monitoring," Nov. 19 1996, uS Patent 5,576,632.

[74] B. A. Whitehead and W. A. Hoyt, "Function approximation approach to anomaly detection in propulsion system test data," *Journal of Propulsion and Power*, vol. 11, no. 5, pp. 1074–1076, 1995.

[75] T. Brotherton and T. Johnson, "Anomaly detection for advanced military aircraft using neural networks," in *Aerospace Conference, 2001,*, vol. 6. IEEE, 2001, pp. 3113–3123.

[76] C. Surace and K. Worden, "A novelty detection method to diagnose damage in structures: an application to an offshore platform," in *Proceedings of the International Offshore and Polar Engineering Conference*, no. 8, 1998, pp. 64–70.

[77] K. Worden, "Structural fault detection using a novelty measure," *Journal of Sound and vibration*, vol. 201, no. 1, pp. 85–101, 1997.

[78] M. Moya, M. Koch, and L. Hostetler, "One-class classifier networks for target recognition applications," Sandia National Labs., Albuquerque, NM (United States), Tech. Rep., 1993.

[79] U. C. Singh and P. A. Kollman, "An approach to computing electrostatic charges for molecules," *Journal of Computational Chemistry*, vol. 5, no. 2, pp. 129–145, 2004.

[80] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[81] L. Jack and A. Nandi, "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms," *Mechanical systems and signal processing*, vol. 16, no. 2, pp. 373–390, 2002.

[82] J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics*, vol. 19, no. 13, pp. 1650–1655, 2003.

[83] H.-E. Kim, A. C. Tan, J. Mathew, and B.-K. Choi, "Bearing fault prognosis based on health state probability estimation," *Expert Systems with Applications*, 2011.

[84] A. Widodo and B.-S. Yang, "Machine health prognostics using survival probability and support vector machine," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8430–8437, 2011.

[85] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," in *Acoustics, Speech, and Signal Processing (ICASSP), on International Conference*, vol. 2. IEEE, 2002, pp. II–1313.

[86] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, on IEEE Transactions*, vol. 32, no. 4, pp. 440–448, 2002.

[87] T. Yairi, Y. Kato, and K. Hori, "Fault detection by mining association rules from house-keeping data," in *Proc. of International Symposium on Artificial Intelligence, Robotics and Automation in Space*, vol. 3, no. 9. Citeseer, 2001.

[88] J. Branch, B. Szymanski, C. Giannella, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," in *Distributed Computing Systems, 2006. ICDCS 2006. on 26th IEEE International Conference.* IEEE, 2006, pp. 51–51.

[89] C. C. Aggarwal, "On abnormality detection in spuriously populated data streams," in *Proceedings of the 5th SIAM Data Min. Conference*, 2005, pp. 80–91.

[90] Q.-B. Gao and Z.-Z. Wang, "Center-based nearest neighbor classifier," *Pattern Recognition*, vol. 40, no. 1, pp. 346–349, 2007.

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

[92] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *Computational Intelligence and Data Mining, 2007. CIDM 2007. on IEEE Symposium.* IEEE, 2007, pp. 504–515.

[93] S. Byers and A. E. Raftery, "Nearest-neighbor clutter removal for estimating features in spatial point processes," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 577–584, 1998.

[94] K. Zhang, S. Shi, H. Gao, and J. Li, "Unsupervised outlier detection in sensor networks using aggregation tree," *Advanced Data Mining and Applications*, pp. 158–169, 2007.

[95] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006, pp. 187–198.

[96] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[97] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[98] F. J. Anscombe, "Rejection of outliers," *Technometrics*, vol. 2, no. 2, pp. 123–146, 1960.

[99] G. Manson, S. Pierce, and K. Worden, "On the long-term stability of normal condition for damage detection in a composite panel," *Key Engineering Materials*, vol. 204, pp. 359–370, 2001.

[100] G. Manson, S. G. Pierce, K. Worden, T. Monnier, P. Guy, and K. Atherton, "Long-term stability of normal condition data for novelty detection," in *spie the international society for optical engineering*. International Society for Optical Engineering; 1999, 2000, pp. 323–334.

[101] A. R. Webb, *Statistical pattern recognition*. Wiley, 2003.

[102] D. Hong, G. Xiuwen, and Y. Shuzi, "An approach to state recognition and knowledge-based diagnosis for engines," *Mechanical Systems and Signal Processing*, vol. 5, no. 4, pp. 257–266, 1991.

[103] J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S. Miksch, and B. Kavsek, "Informal identification of outliers in medical data," in *Proceedings of the 5th*

International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. Citeseer, 2000, pp. 20–24.

[104] M. Desforges, P. Jacob, and J. 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.

[105] W. A. Shewhart, "Economic control of quality of manufactured product." 1931.

[106] C. Chow, "On optimum recognition error and reject tradeoff," *Information Theory, on IEEE Transactions*, vol. 16, no. 1, pp. 41–46, 1970.

[107] L. K. Hansen, C. Liisberg, and P. Salamon, "The error-reject tradeoff," *Open Systems & Information Dynamics*, vol. 4, no. 2, pp. 159–184, 1997.

[108] G. Fumera, F. Roli, and G. Giacinto, "Reject option with multiple thresholds," *Pattern recognition*, vol. 33, no. 12, pp. 2099–2101, 2000.

[109] L. P. Cordella, C. De Stefano, F. Tortorella, and M. Vento, "A method for improving classification reliability of multilayer perceptrons," *Neural Networks, IEEE Transactions*, vol. 6, no. 5, pp. 1140–1147, 1995.

[110] K. Pearson, "Contributions to the mathematical theory of evolution. skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London. A*, vol. 186, pp. 343–414, 1895.

[111] 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.

[112] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.

[113] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.

[114] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[115] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Probability theory and related fields*, vol. 57, no. 4, pp. 453–476, 1981.

[116] H. Motulsky, *Intuitive biostatistics: A nonmathematical guide to statistical thinking*. Oxford University Press, 2010.

[117] W. Staszewski, I. Al-Shidhani, and S. Beck, "Leak monitoring in pipeline networks using wavelet analysis," *Key Engineering Materials*, vol. 245, pp. 51–58, 2003.

[118] G. Sabnavis, R. G. Kirk, M. Kasarda, and D. Quinn, "Cracked shaft detection and diagnostics: a literature review," *Shock and Vibration Digest*, vol. 36, no. 4, p. 287, 2004.

[119] H. C. Pusey and M. Roemer, "An assessment of turbomachinery condition monitoring and failure prognosis technology," *Shock and Vibration Digest*, vol. 31, pp. 365–371, 1999.

[120] S. Nandi and H. A. Toliyat, "Condition monitoring and fault diagnosis of electrical machines-a review," in *Industry Applications Conference, 1999. Thirty-Fourth IAS Annual Meeting. Conference Record of the 1999 IEEE*, vol. 1. IEEE, 1999, pp. 197–204.

[121] L. Pau, *Failure diagnosis and performance monitoring*, ser. Control and systems theory. M. Dekker, 1981.

[122] S. W. Doebling, C. R. Farrar, M. B. Prime, and D. W. Shevitz, "Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review," Los Alamos National Lab., NM (United States), Tech. Rep., 1996.

[123] H. Sohn, C. Farrar, F. Hemez, D. Shunk, D. Stinemates, and B. Nadler, "A review of structural health monitoring literature : 1996-2001," Los Alamos National Lab., NM (United States), Tech. Rep., 2003.

[124] K. Worden, S. G. Pierce, G. Manson, W. Philp, W. J. Staszewski, and B. Culshaw, "Detection of defects in composite plates using lamb waves and novelty detection," *International Journal of Systems Science*, vol. 31, no. 11, pp. 1397–1409, 2000.

[125] M. J. Chae and D. M. Abraham, "Neuro-fuzzy approaches for sanitary sewer pipeline condition assessment," *Journal of Computing in Civil engineering*, vol. 15, no. 1, pp. 4–14, 2001.

[126] W. M. Solano, "Measuring anomaly with algorithmic entropy," 2007.

[127] R. Hartley, "Transmission of information1," 1928.

[128] C. E. Shannon and W. Weaver, "A mathematical theory of communication," 1948.

[129] S. Ihara, *Information theory for continuous system*. World Scientific Publishing Company, 1993, vol. 2.

[130] W. Weaver, "Recent contributions to the mathematical theory of communication," *The mathematical theory of communication*, vol. 1, 1949.

[131] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[132] T. Cover, J. Thomas, J. G. Proakis, M. Salehi, and R. H. Morelos-Zaragoza, *Elements of information theory. telecommunications*. Wiley series, 1991.

[133] F. Nielsen and R. Nock, "On r'enyi and tsallis entropies and divergences for exponential families," *arXiv preprint arXiv:1105.3259*, 2011.

[134] J. Aczél and Z. Daróczy, *On measures of information and their characterizations*, ser. Mathematics in Science and Engineering. Elsevier Science, 1975.

[135] A. Rényi, *Selected papers of Alfréd Rényi*, ser. Selected Papers of Alfréd Rényi. Akadémiai Kiadó, 1976, no. v. 2.

[136] Y. Kopylova, D. A. Buell, C.-T. Huang, and J. Janies, "Mutual information applied to anomaly detection," *Communications and Networks, Journal of*, vol. 10, no. 1, pp. 89–97, 2008.

[137] R. Nowak, *Introduction to Detection Theory.* Online Course, ECE 830, 2011.

[138] J. Neyman and E. S. Pearson, "The testing of statistical hypotheses in relation to probabilities a priori," *Joint Statistical Papers*, pp. 186–202, 1967.

[139] C. Scott, "The neyman-pearson criterion," *Connexions, Jan*, vol. 23, p. 2004, 2004.

[140] J. García-Martín, J. Gómez-Gil, and E. Vázquez-Sánchez, "Non-destructive techniques based on eddy current testing," *Sensors*, vol. 11, no. 3, pp. 2525–2565, 2011.

[141] M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot modeling and control.* John Wiley & Sons New York, 2006.

[142] R. N. MacDonough and A. D. Whalen, *Detection of Signals in Noise.* Access Online via Elsevier, 1995.