DATA COMPRESSION OF MFL SIGNALS

A Graduate Project Report submitted to Manipal University in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING
In

Electronics and Communication Engineering

Submitted by

A.Sriharsha
Reg. No:100907551

Under the guidance of

Debmalya Mukherjee
Scientific Officer
B.A.R.C

&

Pallavi R. Mane
Associate professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY
(A Constituent College of Manipal University)
MANIPAL – 576104, KARNATAKA, INDIA

MAY2014

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

MANIPAL INSTITUTE OF TECHNOLOGY

(A Constituent College of Manipal University)
MANIPAL – 576 104 (KARNATAKA), INDIA

Manipal

CERTIFICATE

This is to certify that the project titled **Data Compression of MFL signals** is a record of the bonafide work done by **A.Sriharsha** (*100907551)* submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Engineering (BE) in **ELECTRONICS AND COMMUNICATION ENGINEERING** of Manipal Institute of Technology Manipal, Karnataka, (A Constituent College of Manipal University), during the academic year 2013-14.

Pallavi R Mane
Project Guide
Asst. Professor
Department of Electronics
And communication
MIT MANIPAL

Prof. Dr. K.PRABHKAR NAYAK
HOD, E&C.
M.I.T, MANIPAL

(*On company letterhead)*

Mumbai
13/5/2014

CERTIFICATE

This is to certify that the project entitled **Data Compression for MFL signals** was carried out by **A.Sriharsha**(*100907551)* at **BHABHA ATOMIC RESEARCH CENTRE, MUMBAI** under my guidance during **January, 2014** to **May, 2014**.

Debmalya Mukherjee
Scientific officer,
Bhabha Atomic research centre,
Mumbai

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to all those people who gave me the possibility to complete this report. Without their persistent support it would not have been possible to carry on this project. I would like to thank the following people.

Dr. Siddhartha Mukhopadhyay, for providing me a project to exemplify the field of data compression. Moreover for providing valuable technical and moral support in all aspects of the project.

My guides **Shri Debmalya Mukherjee (**scientific officer, Control and Instrumentation Division**)** and **Mrs Pallavi R Mane (**Asst. Professor, Department of ECE, MIT Manipal**)** for their exemplary guidance, monitoring and constant encouragement throughout the course of training. The help and guidance given by them time to time shall carry me a long way in the journey of my life of which I am about to embark.

H.O.D of Department of Electronics and Communication Engineering **Dr. Prabhakar Nayak** and Head of Control and Instrumentation Division **Dr. Y. K. Taly** for approving my request to extend my project timeline.

I am obliged to staff members of Department of Control and Instrumentation Division, for the valuable information provided by them in their respective fields. I am grateful for their co-operation during the period of my assignment.

Lastly I Thank almighty, my parents, and my friends for their constant encouragement, without which this assignment would not be possible.

ABSTRACT

Pipelines are economical and environmentally safe mode of transportation of petroleum products over long distances in bulk quantities. Pipeline installations involve high capital investment and are important from strategic point of view. These pipelines are buried below the ground surface and are installed for providing service for a period above 50 years. Although protected by right-of-way surveys, cathodic protection surveys, leak detection programs, excavation to look for pipeline corrosion or protective coating failures,, it is mandatory to do online inspection of the pipelines at a regular interval, by **in-line inspection tools popularly known as in-line inspection tools popularly known as Instrumented pig (IPIG),** leading to yearly inspection of several kilometres of pipelines. Combinations of these procedures constitute an overall integrity assurance program of the pipeline operator. Preventive maintenance based on the inspection report avoids accidental loss of highly inflammable and costly petroleum products. In this regards the area of inline pipeline health has become highly important

The report presents a novel three stage algorithm for online compression of magnetic flux leakage (MFL) signals that are acquired in inspection of oil and gas pipelines. In the first stage, blocks of MFL signal are screened for useful information using a semi-robust statistical measure. Mean Absolute Deviation (µAD). The study presents guidelines for selecting a block size to deliver robust screening and efficient compression ratios. In second stage, a multivariate approach is used to compress the data across sensors using principal component analysis (PCA). The second stage is invoked only when an anomaly is detected by sufficiently large number of sensors. In the third stage, the signal is further compressed within each sensor (univariate approach) using Discrete Wavelet Transform (DWT). Implementation on real-time MFL signals demonstrates the algorithm's ability to achieve high compression ratios with low Normalized Mean Square Error (NMSE) while being fairly robust to baseline shifts.

Successfully implemented the screening algorithm and Principal component analysis on the sample data taken from a binary file and compressed it ten times to the original binary file. The binary file consists of real time data taken from pipeline inspection gauge. Using screening algorithm, the data is successfully reduced. Next implemented the algorithm on the binary file and trying to reconstruct the data obtained from the compressed binary file. Reconstructed the image but could recover only welds.

Many compression algorithms exist both lossy and lossless but still this algorithm remains best out of all. Since here the job is to detect only anomalies i.e., only useful information of the raw data. A pipeline stretches from 600Kms to 900Kms. So a large amount of data will be collected which will take more processing time so this algorithm would be helpful in decreasing the amount of data as much as possible..

LIST OF FIGURES

Contents

CHAPTER 1
INTRODUCTION

1.1 *Introduction*

Identification of a system using multi-sensor data (multi-sensor data fusion) is the integration process where there is an actual combination of sensory information from different sources into one representation format [1]. It deals with the synergistic combination of information made available by various sources in order to provide a better understanding of the scene. Data fusion techniques combine data from multiple sensors and related information from associated databases, to achieve improved accuracies and more specific inferences than could be achieved by the use of a every single sensor alone[2]. The fused data not only reflects the information collected by every source (redundant data), but also provides an insight into information that cannot be inferred by looking at data from each source separately (complimentary data).

The concept of data fusion is hardly new. Living species are known to fuse their sensory organs for better interpretations of its surroundings. For example, it may not be possible to assess the quality of an edible substance based solely on the sense of vision or touch, but evaluation of edibility may be achieved using a combination of sight, touch, smell, and taste. A much needed impetus into the development of data fusion techniques came with the emergence of modern sensors, advanced signal processing and data handling tools, coupled with high speed computing machines. Currently, data fusion systems are used extensively for missile target tracking, automated identification of targets, and limited automated reasoning applications [3]. Nonmilitary applications include monitoring of manufacturing processes, condition based maintenance of complex machinery, robotics, medical applications and data interpretations. In context of this work, it is proposed to fuse data from multiple sensors measuring magnetic leakage flux from magnetized pipeline walls for defect characterization in/on its surface.

Magnetic flux leakage (MFL) technique is one of the most widely used nondestructive evaluation (NDE) methods for pipeline inspection. Petroleum products are transported to consumer sites through a vast network of pipelines. In order to ensure integrity of the system, the pipelines are periodically examined using various NDE techniques, MFL being one of them. An instrumented pipeline inspection gauge (IPIG), used for MFL inspection, consists of a strong permanent magnet that magnetizes a segment of the carbon steel pipeline to near saturation. An array of circumferentially oriented Hall sensors located at the magnetic neutral plane of permanent magnet assembly senses the axial and/or radial component of the leakage flux [4]. Any abrupt change in the thickness of the pipe wall, caused by corrosions or mechanical damage, results in a redistribution of magnetic field in the vicinity of the flaw, causing a change in the pattern of the magnetic flux leakage. This change in leakage flux is sensed by circumferentially disposed array of Hall sensors [5] [6]. Detecting and characterizing defects constitute the overall goal of the MFL inspection procedure.

Carbon steel pipelines are widely deployed in many countries to transport oil and gas products across several thousands of kilometres. Owing to the large scale layout, even a small leakage in the pipelines results in large economic losses. More importantly, since stretches of these of these pipelines carrying inflammable products pass through highly populated areas, poor health of pipelines can give rise to safety concerns. Regular condition monitoring of pipelines is therefore necessary to ensure both public safety and proper transportation of products without loss of economy. MFL is a magnetic method of non-destructive testing that is widely used for both detection and characterization of metal loss defects using IPIG [1]. The IPIG consists of magnetic assembly, data storage and power modules. It travels by the pressure exerted by the flow of product that is being transported in the pipeline. The magnetic assembly consists of an array of permanent magnets and hall sensors.

The technique of MFL testing consists of

(i) Local magnetization of the pipe wall to near saturation and
(ii) Recording flux leakage data in high end digital signal processors.

It is a common practice to magnetize a pipeline axially. Consequently, defects oriented in a way that oppose magnetic flux (e.g. circumferential defects) are detected with greater ease than those oriented otherwise (for e.g. longitudinal defects). However if the width of the defect is sufficiently large, even axial magnetization can detect and characterize longitudinally disposed defects. A typical MFL signal in the absence and presence of metal loss is shown in fig. 1. Ideally the recorded signal should stay constant in the absence of any anomalies. The presence of measurement noise, however, introduces fluctuations as shown in fig 1(a). Such segments of data do not carry any useful information and hence are termed as noisy as noisy blocks.

Interpretation of MFL signals and inversion techniques for defect characterization are discussed in numerous works [2-5]. The focus of this work is on the data acquisition stage of an IPIG operation, specifically the online compression of the large volumes of data that result during this stage. A typical 24’’ in IPIG tool generates 80 GB of data from a single run in the pipeline, which stretches up to 200km. The success of the defect characterization (from data) naturally demands high quality (informative) data while operational constraints do not permit a large capacity storage device. A recent study recommends

(i) Increase in sampling frequency to achieve better characterization of defects.
(ii) Compactness of the storage components, and
(iii) Increase in inspection length of a single run [6].

The foregoing factors combined with a large volume of data that is generated calls for the efficient online data compression algorithm.

The problem of data compression has drawn the serious attention of academia and industry for several decades particularly because they arise in various important applications such as image processing, telecommunications, medicine [7]etc. consequently, sophisticated algorithms for signal compression have emerged, each of them suiting a class of applications. The optimality of the compression algorithm is largely determined by the nature of the signal and the end-use of the signal and end-use of information contained in the signal, both of which being highly dependent on the application. There is hardly a universal algorithm that provides best compression for all types of signals. The literature on the compression of MFL signals is relatively scarce, widely used compression algorithms in other applications serve as suitable candidates for the application under study; yet, there is a need to provide a fresh treatment to the problem for two reasons. Firstly, the signals encountered from the IPIG carry features that differ from those encountered in other arenas. Secondly, the end-use of data is towards detection of corroded areas of pipelines placing our interest only in those parts of signal that capture anomalies. Naturally there is hardly an incentive to store or retain signals that lack any information. Motivated by these reasons, a formal development of an efficient and a simple online compression algorithm is taken up in this work.

The report is about an efficient three stage online algorithm for the compression of MFL signals. The first stage consists of feature detection exercise that is used to screen out uninteresting portions of the signal. The algorithm is based on the statistical measure of variance (excitation) namely, the Mean Absolute Deviation (µAD). At this stage only those portions of the raw MFL data that contain some useful information about the pipeline anomalies are retained, while discarding other blocks of data. This step contributes significantly to the overall compression that is achieved with the proposed algorithm. The effectiveness of this stage depends on the size of the block being screened relative to the axial length of feature.

The second stage of consists of compression across sensors where a multivariable dimensionality reduction technique is employed. For this purpose, The Principal Component Analysis (PCA), a well-established dimensionality reduction technique is applied. The assumption of linear relationships between sensor readings may not be valid, but the objective is not extract relationships. The aim is to deploy an affective multivariable compression tool that can easily implemented online. Among the numerous techniques that suit the needs, PCA is a competitive choice because of its computation simplicity and theoretically efficacy. [8] The idea in PCA is to represent the same information (contained in the raw data) in a lower-order virtual sensor or principal component space. The virtual sensor spaces are orthogonal to each other. The benefit of compression achieved in this stage naturally depends on the number of sensors spanned by the anomaly (sensor span) and the correlation cross those sensors. Therefore, this step is invoked only when the sensor span of the anomaly is large. Such strategy avoids computational burden that is not worth the effort. Consequently when anomaly such as metal loss that usually has a small sensor span, is detected, the compression algorithm retains the raw MFL data, without trans-forming it to principal components. In the third and final stage of compression algorithm, the strategy is to exploit the correlation within a single sensor's reading for compression. A natural choice of technique achieves this task is the wavelet transform [9]. Wavelets have proved to be very effective compression tools in numerous applications such as image compression, biomedical signal compression and process data compression to name a few [10]. A Discrete Wavelet Transform (DWT) using Daubechies wavelets is employed for this purpose. Compressing the signal using DWT also denoises the signal, which improves the characterization of the pipeline anomalies. The success of the overall compression is measured by an appropriate metric, namely, the compression ratio. Implementation of the proposed method achieved in each dataset depends on the extent and type of anomalies present in the data. A dataset with fewer anomalies will yield a larger compression ratio.

1.2 *Motivation*

Instrumented pigs work on either magnetic flux leakage (MFL) or Ultra – sonic (UT) principle. The main advantage of MFL over UT is principle is that the former requires no coupling medium for sensing and hence can be used for both liquid and gas pipelines. Magnetic Flux Leakage (MFL) technique is one of the oldest and most commonly used technique for detecting corrosion in the pipe wall as well as pipeline features like welds, valves etc. How to analyse and compress the MFL data is? This project best explains how to analyse the data. Both multivariate analysis and univariate analysis are involved in this project. In multivariate analysis why we are preferring to use PCA rather than linear discriminant analysis (LDA) and Independent component analysis (ICA). So far the results of the project are giving satisfying results.

1.3 *Objective of the work*

Implement a three stage compression algorithm on real time data collected by IPIG. Analyze how much data is compressed compared to the original and how much it is deviating from the original data.

1.4 *Target Specifications*

Our aim is to implement the compression algorithm on a file containing real time data of Pipeline features and defects. The algorithm should be implemented on the file without losing any information and achieve as much compression as possible.

1.5 *Project Work schedule*

- *January 2014*
  - Data screening Of MFL signal to differentiate between a noisy block and an informative block using a screening algorithm.
  - Finding the threshold for the screening algorithm.

- *February 2014*
  - Multivariate compression using Principal Component Analysis.

- *March 2014*
  - Implement the algorithm of mean absolute deviation on entire file system

- *April 2014*
  - Implement the Screening algorithm and Principal Component Analysis on whole data

- *May 2014*
  - Documentation

1.6 *Organization of the Project Report*

The report has been organized into 5 chapters with first introducing IPIG and how the mechanism works and what kind of analysis has to be done to detect the signal. In The second chapter IPIG module has been explained in detail. In the Third chapter the Algorithm has been explained why we are using multivariate and after that univariate can be understand clearly from the explanation. In the Fourth chapter the results from the image compression are attached. In the fifth chapter conclusions were drawn and future work has been reported.

CHAPTER 2
BACKGROUND THEORY AND LITERARY REVIEW

2.1 *Introduction*

Magnetic flux leakage techniques were used as early as 1868 by the Institute of Naval Architectures in England, where defects in magnetized cannon tubes were found with compass [12]. With the developments in MFL technology since then, it became a cost-effective, non-destructive alternative to hydrostatic testing, for pipeline integrity assessment [11]. Early 20th century witnessed a tremendous growth in the demands of petroleum products, which resulted in lying of buried cross-country pipelines all across the globe. With time, in line inspection (ILI) and maintenance became a critical issue for the pipeline industry. Even before the NDE of pipelines started, operational personnel realized that contaminate were blocking their pipelines and used bundles of rags tied with baling wire to clear them. In late 1940s companies were becoming extremely conscious of internal problems associated with metal fatigue and corrosion. “Shell research “proposed (1963) and successfully developed the first IPIG, which sought out areas of corrosion and recorded the defects internally on film. The first and only run of this pig was performed on a 10-inch pipeline in Texas. The method utilized to detect corrosive areas was eddy current technology, which was later dropped in favour of flux leakage technology, which in turn was the forerunner of the many pigs currently being run today [13].

An IPIG contains a strong permanent magnet that magnetizes the ferromagnetic pipe-wall to near saturation. An array a circumferentially disposed Hall–effect sensors, located in the midpoint of the north and South Pole of the permanent magnet assembly senses the leakage flux from the pipe-wall. When the PIG passes through a defective pipeline region, some of the magnetic flux leaks into the surrounding and its axial/radial component are measured using the circumferential array of Hall-effect sensors. The measured using the circumferential array of Hall-effect sensors. The measured using the circumferential array of Hall-effect sensors. The measured Hall voltage is proportional to the leakage flux density and constitutes the MFL signal. In a defect free region, there is no leakage and the low uniform amplitude signal is measured. In the presence of metal loss corrosion the flux leaks into the surrounding producing a local change in the measured MFL signal. The information in the signal is utilized to detect and characterize anomalies in the pipe wall.

In this chapter we will discuss the technique how we are going to capture the signals and how it will help in analysing the health of pipeline, Instruments of IPIG, Technology Development and overview of compression techniques. Various compression algorithms will be discussed and will analyse how they fall short for the problem.

2.2 *Technique*

A section of pipe is saturated with strong rare earth (NdFeB) permanent magnets. In case of break in pipe geometry (internal or external) due to metal loss level of leakage flux changes near the wall. The magnetic flux leakage is sensed by two sets of hall sensors. The second sets of sensors are used for differentiating internal and external defects. The measured leakage field depends on the radial depth, axial length, circumferential width and shape of the anomaly, as well as the magnetic properties of the nearby material. As the instrument moves along the pipeline propelled by the product flowing in the pipeline, the hall sensors sense the leakage flux density continuously and the outputs are acquired, digitized and in the on-board data acquisition and storage system of the instrument.

2.2.1The Instument

The instrument pig consists of magnetic module, data acquistion system module (DAS), battery module and pig locator module. The polyurethane cup mounted on IPIG,seals the pipe and the pressure of oil flowing through these pipelines gives required propelling force for its movement. Front cups are sealing cups and other cups are supporting cups. It also consists of 3 odometers to record distance travelled by the instrument inside the pipeline. An inclinometer circuit is used to measure the inclination of IPIG during its travel in the pipeline. This information is useful in locating clock position of the faults.

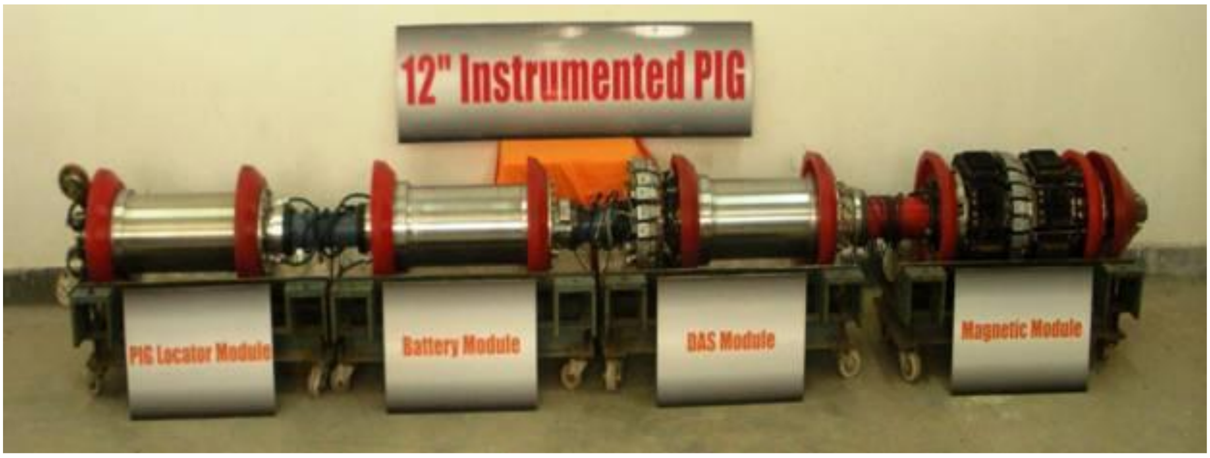


Figure 2.1. Instrumented Pipeline Gauge

2.2.2Magnetic Module

The magnetic module has eight or six segments upon design of IPIG that covers almost the total circumference of the pipe. Each segment consists of two strong permanent Neodymium Iron Boron (NdFeB) magnets placed on two ends of the backing iron which magnetize the pipeline along axial direction. Brushes have been provided over the magnets which have direct contact with the instrument surface of pipe wall and offer low reluctance path to the flux entering the pipe wall. Spring loaded polyurethane sensor arms fitted with hall sensors are mounted on the central portion of the backing iron in the form of a ring. The magnetic module is designed to negotiate the weld protrusions, 5-D bends, tees & other fitting etc.

2.2.3 DAS Module

The multiprocessor data acquisition system has been used to acquire data sensed and transmitted by the hall probe sensors. The data is stored to solid state memory which can withstand high level of shocks and vibration. The electronics are housed in a pressure tight vessel on anti-vibration mountings. Data from each hall sensor is acquired after every 2 mm travel of IPIG through the pipeline. Vibration, temperature and rotation data is also acquired from respective sensors during the travel of IPIG.

2.2.4 Battery Module

This module houses the power supply system for the IPIG. Presently it has two battery packs for powering the sensors and electronic cards of DAS module. The battery packs consist of series of high capacity non chargeable lithium cells. Each battery pack is connected to a DC-to-DC converter, which supply power to DAS module. The battery module is capable to inspect approximately 300 KM in one inspection run.

2.2.5 Pig Locator Module

This module consists of a transmitter capable of continuously emitting low frequency electromagnetic signals. The signals are picked up by an antenna and receiver kept above ground within a range of 10-20 meter diameter from the transmitting signal. This module helps in tracking the movement of IPIG during its travel. Odometer assembly having three odometer wheels is fitted in the rear side of pig locator module. These odometers are used to record the axial movement of IPIG for location identification.

2.3 Technology Development

2.3.1 Magnetic circuit development and analysis

Why NdFeB?

Neodymium iron boron (Nd-Fe-B) magnets have a higher Maximum Energy Product, (BH) max, than Sm-Co magnets.

- (BH) max of Nd-Fe-B can easily reach 30 MGOe and even goes up to 48 MGOe.
- Generally, the cost of Sm-Co magnets is higher than Nd-Fe-B magnets.
- MFL based IPIG work in environment where temperature can go up to 50 deg. Celsius

2.3.2 FEM Analysis

FEM methods have been widely used to study various types of problems related to IPIG. These include study of effects of eddy currents, magneto-mechanical analysis, inverse problem solution etc. Leakage profile due to MFL phenomenon in IPIG can be obtained by solving nonlinear Poisson equation. Where, A is magnetic vector potential, is material nonlinear reluctivity and J is imposed current density. Defect width along Z direction (into the paper) affects leakage flux. Hence, in any attempt to compare leakage value given by FEM with those obtained practically, one has to go for 3D modelling and analysis. Model of one sector of magnetic module was solved as magneto static nonlinear problem using Lagrange quadratic elements. The peak value of leakage flux is affected by length and depth of defect. When IPIG moves in pipeline, operational variables such as PIG velocity and fluid pressure also affect leakage profiles. Faster the instrument moves through pipe, more are the eddy currents opposes the field due to IPIG. Hence effective field and consequently leakage flux for a given defect decreases. This problem was simulated as nonlinear transient problem.

2.3.3 Data analysis technique and in-house software development

The evaluation involves analysing large volume of data from an inspection with stipulated accuracy within a limited time-schedule. The task can only be handled with very high level of automation, reducing the time for offline assessment of data as well as increasing the reliability of the same. Prior to storage, the data collected by IPIG is processed on-line by thresholding its projections on a set of wavelet basis, to retain useful information regarding pipe features and metal loss defects. The compressed MFL data is decompressed and de-noised off line using discrete wavelet transform (DWT) to form an image of the pipe surface. Pipe features and defects are detected from the pipe image using image segmentation technique. The three primary signal features detected are axial extent of signal termed as span, circumferential spread of signal in terms of number of sensors and maximum peak to peak gauss level for particular feature. In addition to these parameters the secondary parameters like shape of the circumferential flux pattern and its spread are also considered. These parameters are then used in the classifier module to finally predict the defect feature dimensions namely length, width and depth. Pre-processing of raw data In off-line processing, the raw MFL data from a run is first scanned for preliminary information on the quality, continuality, environment (ambient temperature, vibration level etc.) and duration of data stored to ascertain the healthiness of the run. It also involves correcting signals with reference to calibration measurements carried out for the sensors response from the data collected on a full periphery groove of uniform wall loss.

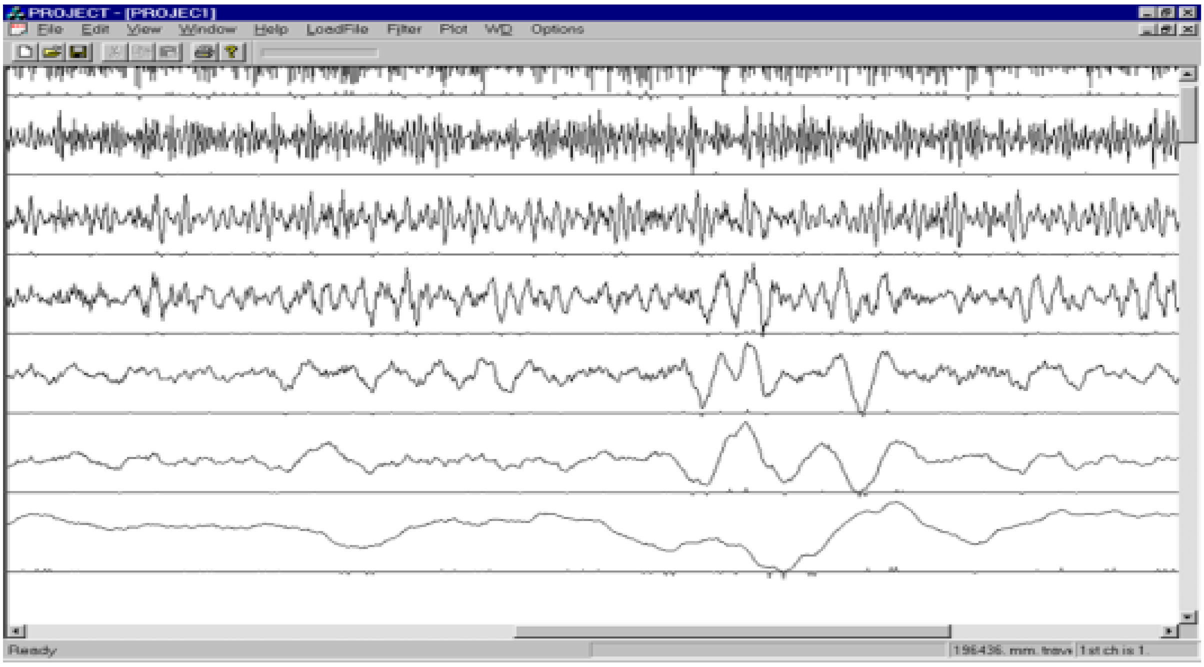


Figure 2.2 Raw MFL signal and the signal decomposed in Time Scale

2.3.4 Characterization or sizing of defects

We use signal from radial sensors to extract defect parameters. The radial MFL signal for a defect is bipolar and is characterized by three parameters-

- Peak to peak span

- Number of sensors and
- Maximal peak to peak flux density Bmax.

The difference in length unit between positive and negative peak called span indicates the length of a defect. Width is estimated unit between positive and negative peak called span indicates the length of a defect. Width is estimated as an empirical function of number of sensors N, sensor pitch p and maximum peak to peak leakage flux density Bmax. It has experimentally found, %WL is a function of Bmax, the ratio of estimates of width and length and the area of the metal loss defect. The estimated values of length and width and the measured value of Bmax are used to calculate %WL using an empirical relationship. The maximum value of peak to peak leakage flux density under the defect is found to be a higher order nonlinear function of % WL, length and width.

An unsupervised defect characterization algorithm works on the segmented image and automatically generates a report giving details of the size and shape of the defect. This package along with an automatic feature and defect characterization package is used for analysis of large volume of data and generation of report.

2.4 Locating defects and pipe features and reporting

The first step of data analysis is to locate magnetic markers placed approximately every one or two kilometres along the pipe. These markers are placed to correct the error in distance measurement due to slippage of odometers. Other pipe features are located with reference to magnetic markers. The distance error due to odometer slippage, is compensated at each weld so as to get zero error at each marker and thus absolute weld log distance is found. Nearest weld log distances (preceding and succeeding) are used to locate any other feature placed between them. Finally a table is formed having all detected pipe features with its log distance from nearest magnetic marker. The algorithm to detect features is mostly rule based and is validated by matching with pipe features from a number of field trials in actual oil pipelines.

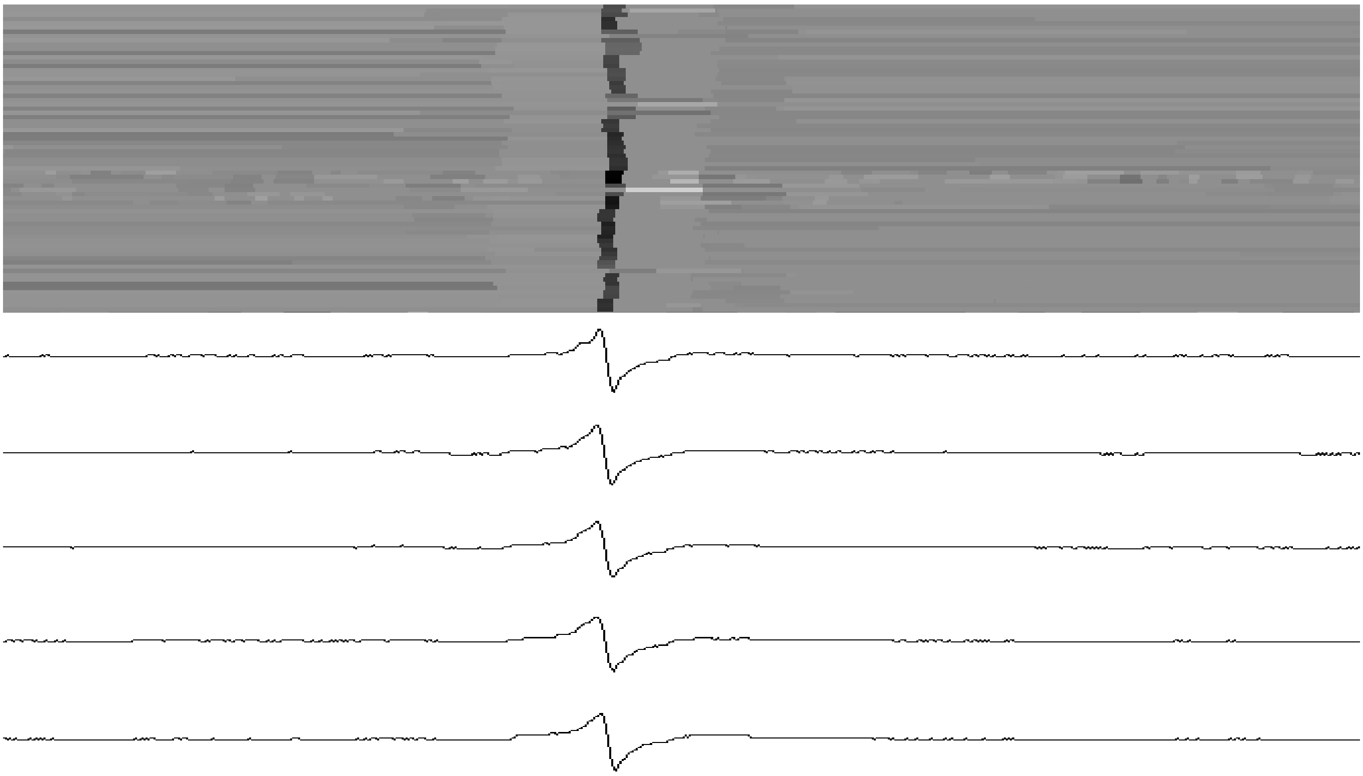


Figure 2.3 MFL signature of weld

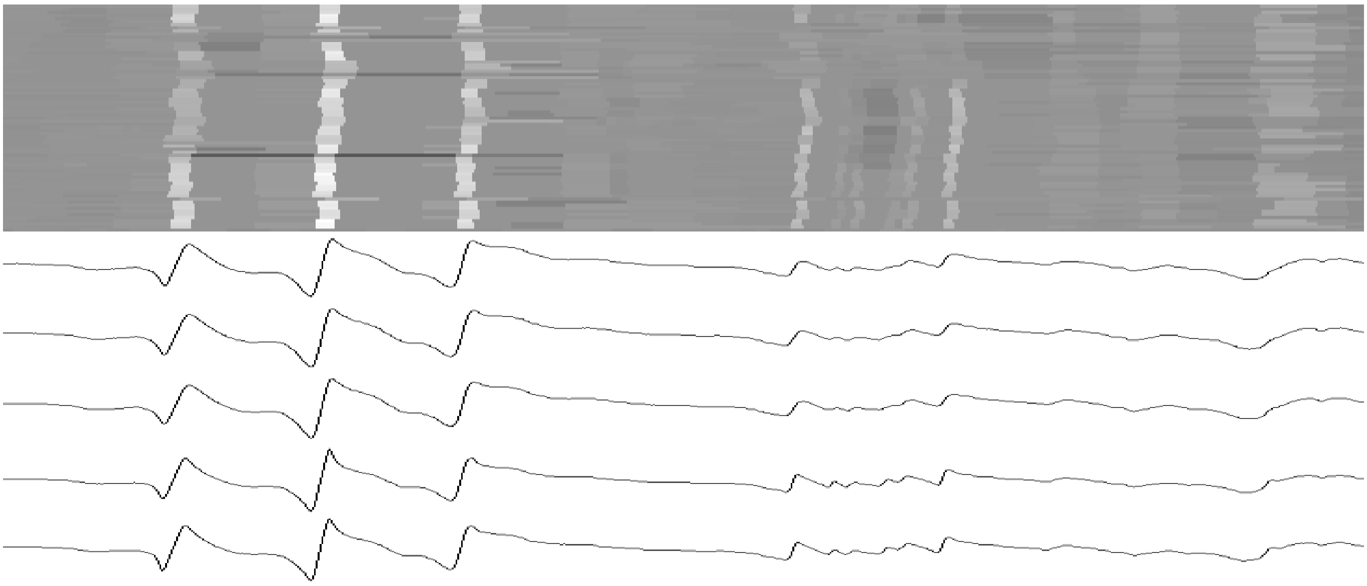


Figure 2.4 MFL signature of valve

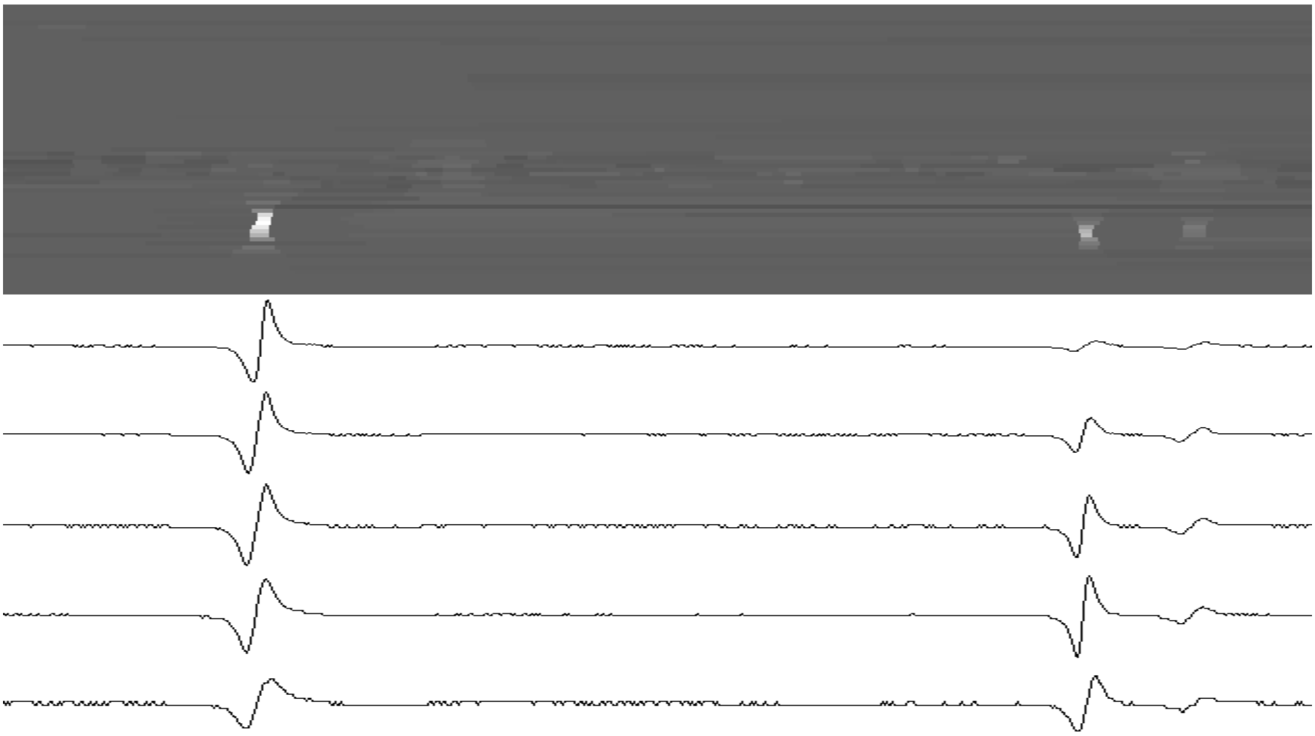


Fig 2.5 MFL signature of Corrosion defects

2.5 *Overview of compression Techniques*

There are two kinds of compression techniques based on their ability to exactly reconstruct the original data, namely, the lossless and lossy compression techniques. The performance of any compression algorithm can be evaluated using two indices-Compressions Ratio (CR) and Normalized Mean Square Error (NMSE), defined as

$$CR = \frac{\text{No. of original samples}}{\text{No. of retained samples}} = \frac{N}{R}$$

$$NMSE = \frac{\sum^{N}_{n=1}(x[n] - x'\,[n])^{2}}{\sum^{N}_{n=1}(x[n])^{2}} \qquad (1)$$

Where x[n] is the original signal. X' [n] is the estimated signal (i.e., the decomposed signal), N is the length of X[n] and R is the length of the compressed signal.

The basic difference between the lossless and the lossy compression techniques is the trade-off that they provide between the NMSE and the compression ratios that can be achieved. Lossless techniques insist on zero reconstruction error whereas lossy techniques are willing to sacrifice the zero reconstruction error property to achieve much higher compression ratios. Given the incentive, lossy techniques are popularly deployed with implicit understanding that the reconstruction error is within pre-defined tolerance levels. The lossy compression techniques are broadly classified into four categories below.

Direct Data Compression Techniques (DDCT) exploit correlation among successive samples to reduce redundancy. The turning point (TP) algorithm replaces every three data samples with two using that best represent the slope. The compression ratio achieved using TP is usually limited two. The amplitude Zone Time Epoch Coding (AZTEC) converts the signal into horizontal lines (plateaus). The reconstructed signal has discontinuities and distortion with large NMSE. The Coordinate Reduction Time Encoding (CORTES) is a combination of both TP and AZTEC. In CORTES the choice of either TP or AZTEC depend upon the nature of underlying signal.

Model-Based Compression Techniques (MCT) rely on a model that represents the signal. Instead of the original data, the model parameters are stored. During the reconstruction stage, each data sample is predicted or interpolated using the model parameters. Vector Quantization (VQ) maps a set of vectors into predefined vector set in the codebook. Parameter extraction-based compression Techniques (PCT) detects and preserves only the required properties of the signal.

In Transform-based Compression Techniques (TCT), the idea is to transform the signal into a domain which facilitates signal representation in terms of very few coefficients. For example, a sine wave typically requires large number of samples to represent it in time domain. However, it is most compactly represented in terms of merely three parameters- the amplitude phase and frequency. The Fourier transform of the signal represent achieves this representation since it uses sines as basis functions. Walsh Hadamard Transform (WHT) is one of the simple and fast transform to be implemented. WHT is unitary and orthogonal transform to be implemented. WHT is unitary and orthogonal transform composed by rectangular waveforms with values +1 and -1. The discrete Fourier Transform (DFT) projects the signal onto set of orthogonal sine and cosine basis functions [14]. Discrete Cosine Transform (DCT) is also similar to DFT except that the cosine wave is basis function. For narrowband signals, Good compression can be achieved using DFT, DCT and WHT since basis function have similar properties. Several real-life applications generate signals that have time-varying frequency content. The aforementioned techniques are not ideally suited for compression of such signals. Of the several extensions that exist for handling time-frequency content, the Discrete Wavelet Transform (DWT) stands out as an excellent tool for compression. The DWT essentially represents the signal in terms of projections onto set of non-redundant basis functions that have (near) compact spread in the time-frequency plane [10]. In fact, DWT is also suited for compressing broadband signals [12].

Linear discriminant analysis (LDA) and the related Fisher’s linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and categorical dependent variable, whereas discriminant analysis has continuous independent variables and categorical dependent variable (i.e. the class label). Logistic regression and probit regression are similar to LDA, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any differences in class and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also criterion variables) must be made.

In statistics, canonical-correlation analysis (CCA) is a way of making sense of cross-covariance matrices. If we have two vectors X= (X1 ,…..,Xn) and Y = (Y1 , ………, Ym) of random variables, and there are correlations among the variables, then canonical-correlation analysis will find linear combinations of the X_i and Y_i which have maximum correlation with each other. This method was first introduced by Harold Hotelling.

All foregoing data compression algorithms exploit the correlation among the samples in a single channel. The across-sensor correlation can also be exploited calling for a deployment of multivariate data compression tools. Principal Component Analysis, which was introduced by Pearson as method for analysing data in a lower dimension space emerges as a ubiquitous choice for multivariable data compression. The driving engine for PCA is the Singular Value Decomposition (SVD) (of the data matrix) or the Eigen value decomposition of the sample covariance matrix. PCA has a striking resemblance with Karhunen-Loeve transform (KLT), which works on covariance matrices of jointly stationary multivariable random process.

Dimensionality Reduction is method for reducing variables under consideration and can be divided into Feature selection and feature reduction. Feature selection also known as variable selection is the process of selecting a subset of relevant features for use in model construction. The central assumption is when using feature selection is that data contain many redundant or irrelevant features provide no useful information in any context.

Feature extraction transforms data in the high dimensional space to a space of fewer dimensions. The data transformation may be linear as principal component analysis. But there are many nonlinear dimensionality reduction techniques. For multi dimensionality data, Tensor representation can be used to reduce the data.

The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low dimensional representation is maximized. In practice, the correlation matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few Eigen vectors can often be interpreted in terms of the large scale physical behaviour of the system. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few Eigen vectors.

Factor analysis is a statistical method used to describe variability among the observed correlated variables in terms of a potentially lower number of unobserved variables called factors, for example, it is possible that variations in four observed variables mainly reflect the variations in two unobserved variables. Factor analysis searched for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus error terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables of observed variables. Factor analysis is related to principal component analysis (PCA), but the two are not identical. Latent variable models, including factor analysis, use regression modelling techniques producing error terms, while PCA is a descriptive statistical technique. There has been significant controversy in the field over equivalence

1. It is sometimes suggested that principal component analysis is computationally quicker and requires fewer resources than factor analysis is computationally quicker and requires fewer resources than factor analysis.

2. PCA and factor analysis can produce similar results. But Factor analysis takes into account the random error that is inherent in measurement, whereas PCA fails to do so.

In signal processing, independent component analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents. This is done by assuming that the subcomponents are non-Gaussian signals and they are statistically independent from each other. ICA is a special case of blind source separation. Independent Component Analysis attempts to decompose a multivariate signals. As an example, sound is usually a signal that is composed of the numerical addition, at each time t, of signals from several sources. The question then is whether it is possible to separate these contributing sources from the observed total signal. When the statistical independence assumption is correct, blind ICA separation of a mixed signal gives very good results. It is also used for signals that are not supposed to be generated by a mixing for analysis purposes. An important note to consider is that if N sources are present, at least N observations are needed to recover the original signals. This constitutes the square case (J = D, where D is the input dimension of the data and J is the dimension of the model). Other cases of undetermined (J > D) and over determined (J < D) have been investigated. That the ICA separation of mixed signals gives very good results are based on two assumptions and three effects of mixing source signals. Two assumptions:

1. The source signals are independent of each other.

2. The distributions of the values in each source signals are non-Gaussian.

Three effects of mixing source signals:

1. Independence: As what we assume, the source signals are independent; however, their signal mixtures are not. That is because the signal mixtures share the same source signals.

2. Normality: Based on the Central Limit Theorem, The distribution of a sum of independent random variables tends towards a Gaussian distribution. Loosely speaking, a sum of two independent random variables usually has a distribution that is closer to Gaussian than any of the two original variables. Here we consider the value of each signal as the random variable.

3. Complexity: The temporal complexity of any signal mixture is greater than that of its simplest constituent source signal.

Those principal contributes to the basic establishment of ICA. If the signals are happened to be extract from a set of mixtures are independent like source signals, or have non-Gaussian histograms like source signals, or have low complexity like source signals, then they must be source signals.

Evidently there exists no ideal universal compression algorithm that is suited for all classes of signals. In fact, the choice of any compression technique depends on factors such as

(i) Nature of the underlying signal
(ii) Important parameters to be retained while reconstructing the signal
(iii) The end use of the data.

The compression algorithm presented in this report is developed in light of the aforementioned factors. The remainder of the report is devoted to the development and demonstration of the proposed three stage compression algorithm.

3. Proposed online compression methodology

The three stage algorithm that is developed for the online compression of MFL signals is presented in this section. The implementation of three stage algorithm is schematically shown in fig 6.

3.1 *Stage I: MFL feature detection algorithm*

The idea is to apply a screening algorithm that explores the variability in the MFL signal to differentiate between a noisy block and an informative block. The tool requires for this purpose should be sensitive to variations and robust to baseline shifts. Statistics offers a variety of measures such as variance, Median Absolute Deviation (MAD) and Mean Absolute Deviation (μAD) as effective measures of variability. Among these three measures, MAD offers the maximum robustness to outliers and shifts. For a univariate series X = {x [1], x [2]… x [N]}, median absolute deviation from the data's median.

$$\text{MAD} = \text{median}(x-\text{median}(x)) \qquad (2)$$

The robustness of MAD is attributed to the property of median, which can robustly estimate the average of a sequence in the presence of noise and outliers. For the application in hand, MAD is not a desirable candidate for screening since it is not sensitive to small amounts of deviation present in a signal is constant. The value of Mad is zero, which indicates that there is no deviation in the signal, when more than half part of the signal is constant.

In contrast, the Mean Absolute Deviation, defined as the mean of absolute deviation from the median,

$$\mu\text{AD} = K*\text{mean}(x-\text{median}(x))$$

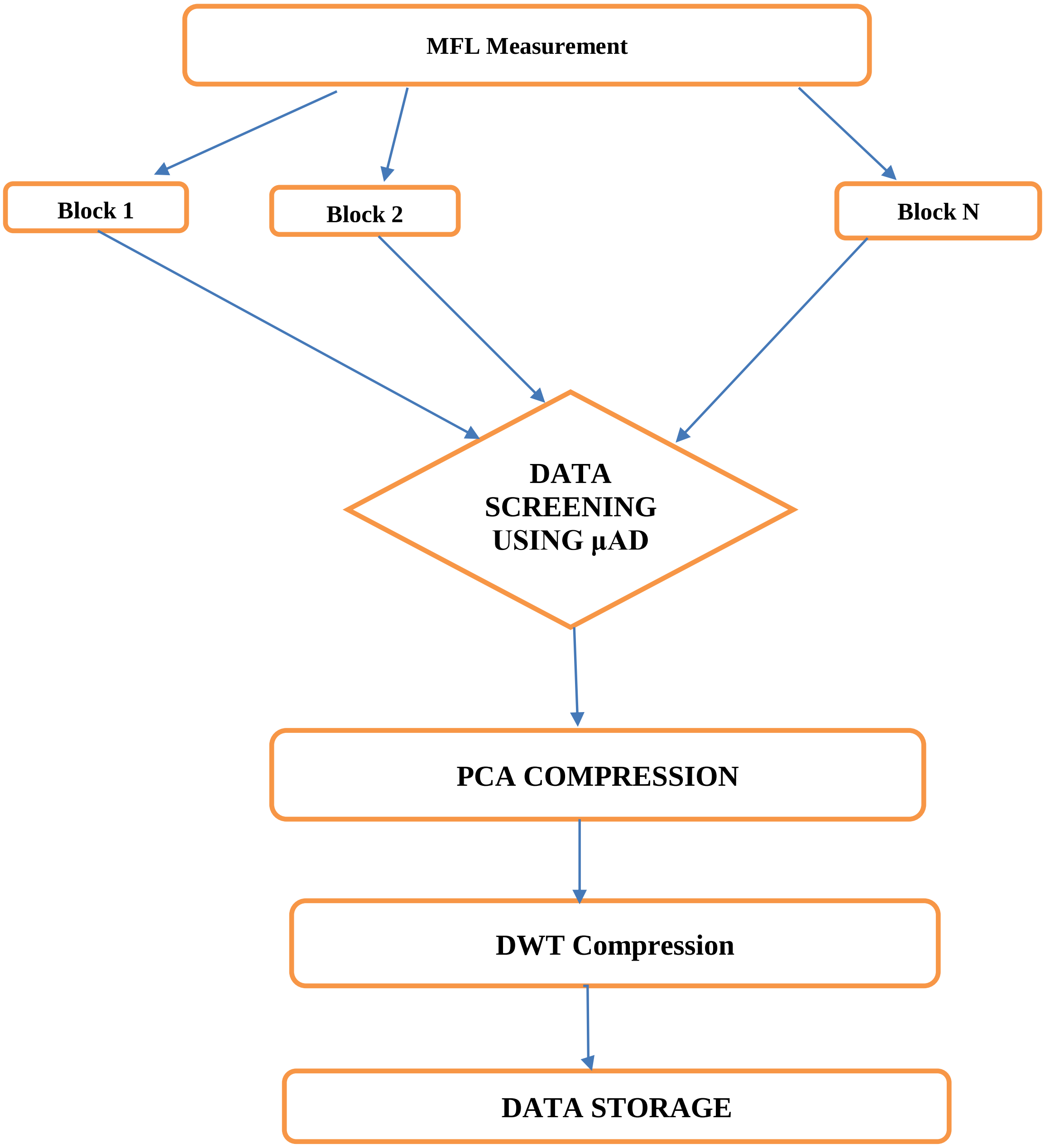


Figure 3 FlowChart for the proposed Compression algorithm

Where k is a scaling factor and depends upon the probability distribution of the noise present in the signal, offers a good mix of robustness and sensitivity that is desires for the application. For Gaussian distribution, the scaling factor is estimated as 1.25 using Monte-Carlo simulations, which also approximately equal to the theoretical one [11].

Steps involved in the screening algorithm

1. Divide the signal into blocks and calculate µAD for each block.
2. If the µAD value is insignificant, reject that particular data block
3. Else retain the data along with the spatial location.

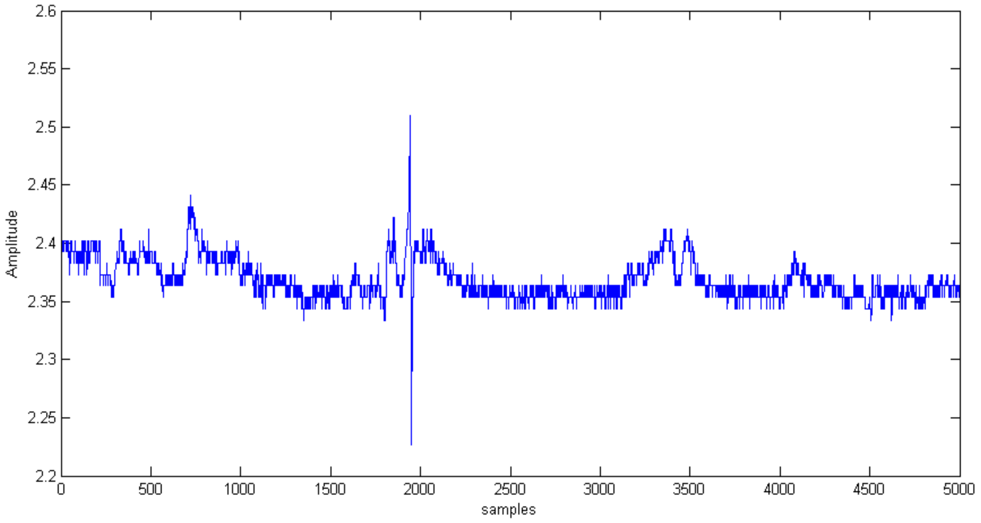


Figure 3.1 Plot of data of a single channel

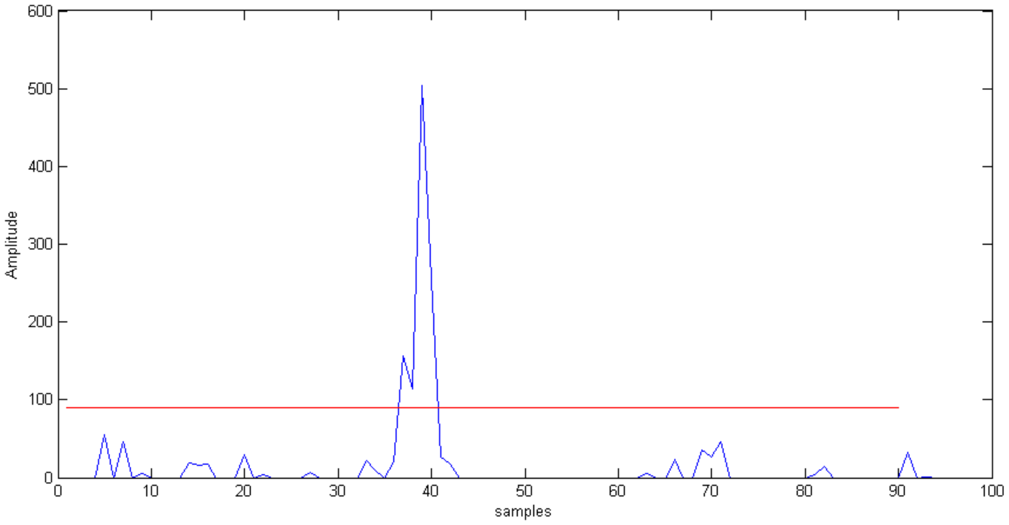


Figure 3.2 Plot of Mean absolute deviation for a single channel with threshold

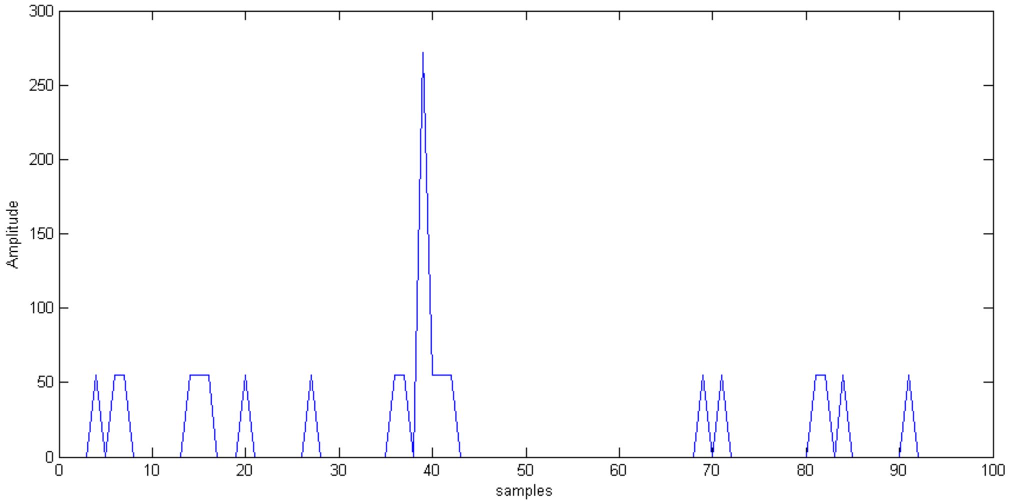


Figure 3.3 Plot of a single channel of Median Absolute Deviation

3.2 Parameters influencing the screening algorithm

The effectiveness of the screening algorithm is significantly influenced by two parameters, namely, the block size and the choice of threshold. The discussion below defines these parameters and presents guidelines for the parameter settings.

Block Size is defined as the number of samples per block. All the block sizes that are further discussed are in the power of two to unable easy computation of the DWT in the third stage. The efficiency of the screening algorithm to detect the presence of any anomaly is highly dependent on the feature to block ratio (FBR) as defined below,

$$FBR = \frac{\text{No. of samples spanning the features}}{\text{Block Size}} = \frac{FS}{BS} * 100$$

Where the signal feature is defined as that part of the signal showing significant leakage. The feature length is not identical to the length of the anomaly/defect; in fact, it is usually longer than that of the defect. In some sense, therefore, it is reflective of the length of the anomaly. For a fixed feature length, FBR decreases as the block size (BS) is increased thereby increasing the possibility of the feature being undetected by the feature detection algorithm. The reason is the increased contribution of noise to the overall variability. Guidelines to choose the right block size can be provided by considering extremes. A very small block size allows accurate detection of the smallest feature but has a reverse impact on the online signal processor. The processor should complete the screening cycle before the next block of data arrives. The complexity of operations involved in the proposed algorithm. An additional risk with a very small block size is that nosy blocks may also be classified as informative blocks. A large block size provides a healthy margin between the screening efficiency of the screening algorithm. Thus, there is trade-off involved.

3.2.1Threshold

The role of threshold, as with several other algorithms, is to practically differentiate between a noisy block and an informative block. Theoretical determination of the threshold is unimaginably difficult and impractical. The natural alternative is to use an empirical approach using field data. The choice of the threshold depends upon the µAD value of the smallest pipeline feature and data block which contains maximum level of noise. A higher threshold value can increase the compression ratio but also increases the probability of a defective region going undetected. Similarity, a smaller threshold will increase the probability of retaining data containing anomalies but will decrease the compression ratio.

3.2.2Robustness

The baseline of the MFL signal may vary due to changes in pipeline thickness, permeability etc. To test the robustness of µAD towards changing baseline, two MFL signals are analysed. The proposed algorithm is bale to judge the blocks containing features as informative despite the presence of baseline shifts. Furthermore, it is rightly classifies the blocks with pure baseline variations as non-informative.

3.3 Multivariate compression using PCA

The Principal Component Analysis is a way of identifying patterns in data, and expressing the data in such a way to highlight their similarities and differences. Since patterns in data hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing the data.

The other main advantages of PCA is that once you have found these patterns in the data, and you compress the data, i.e. By reducing the number of dimensions, without much loss of information. This technique used in image compression

3.3.1 Representation

When using these sorts of matrix techniques in image processing, we must consider representation of images. A square, N by N image can be expressed as an N^2 dimensional vector.

$$X = (\quad x_1 x_2 \; x_3 \; . \; . \; . \; . \; x_{N^2} \quad)$$

Where the rows of pixels in the image are placed one after the other to form a one dimensional image. E.g. The first N elements (x_1-x_N will be first row of the image, the next N elements are the next row, and so on. The values in the vector are the intensity values of the image, possibly a single greyscale value.

3.3.2 PCA to find patterns

Say we have 20 images; each image is N pixels high by N pixels wide. For each image we can create an image vector as described in the representation section. We can then put all the images together in one big image-matrix like this:

Images Matrix = [ImageVec1, ImageVec2......ImageVec20]

Once we have our original data in terms of the Eigen vectors we found from the covariance matrix. Why is this useful? Say we want to do facial recognition, and so our original images were of people faces. Then, the problem is, given a new image, whose face from the original set is it? The way this is done in computer vision is to measure the difference between the new image and original images, but not along the original axes, along the new axes derived from the PCA analysis.

It turns out that these axes works much better for recognising faces, because the PCA analysis has given us the original images in terms of the differences and similarities between them. The PCA analysis has identified the statistical patterns in the data.

Since all the vectors are N^2 dimensional, we will get N^2 eigenvectors. In practice, We are able to leave out some of the less significant eigenvectors, and the recognition still performs well.

3.3.3 PCA for image Compression

Using PCA for image compression also known as Hotelling, or Karhunen and Leove (KL), Transform. If we have 20 images, each with N^2 pixels, we can form N^2 vectors, each with 20 dimensions. Each vector consists of all the intensity values from the same pixel from each picture. This is different from the previous work because before we had a vector for image, and each item in that vector was a different pixel, whereas now we have a vector for each pixel, and each item in the vector is from a different image.

Now we perform the PCA on this set of data, we will get 20 Eigen vectors because each vector is 20-dimensional. To compress the data, we can then choose to transform the data only using, say 15 of the eigenvectors. This gives us a final data set with only 15 dimensions, which has saved ¼ of space. However when the original data is produced, the images have lost some of the information. This compression technique is said to be lossy because the decompressed image is not exactly the same as the original, generally worse.

The multivariate compression is invoked only when a particular anomaly is detected by a large number of sensors. For instance, a weld in the pipeline spans across all the sensors. Thus two or three blocks, depending upon the axial length of the weld, will be detected across all the sensors using the feature detection algorithm.

Principal Component Analysis is a linear orthogonal transform of measurement from a p-dimensional space to another p-dimensional space to another p-dimensional space, so that the coordinates of the data in the new space are uncorrelated and the greatest amount of variance of the original data is expressed by only few coordinates. Let p variables X1, X2… Xp be transformed to another p variables pc_1, pc_2, pc_3… pc_p called principal components. The principal components are arranged in the order of decreasing variance.

The variables X_1, X_2… X_p are standardized to have zero mean and unity variance primarily to avoid ill-conditioning. The principal components can be calculated through an Eigen value decomposition of the sample covariance matrix (X^TX) where

$X = [X_1\ X_2\ \ldots\ Xp]$

The ith principal component is calculated as a linear combination of the variables

$PCi = e_{i1}X_1 + e_{i2}X_2 + e_{i3}X_3 + \ldots + e_{ip}X_{ip}$, i=1, 2, 3... P

Where the constants ei1, ei2... eip are the elements of the corresponding eigenvector (also called as loadings) of the covariance matrix [12].

The theoretical number of principal components is equal to number of sensors that have captured the anomaly. The useful information is however contained in a much fewer principal components as indicated by the significant eigenvalues of the covariance matrix. Retaining only the useful ones produces significant compression [13].

3.4 Method for implementation of PCA analysis

3.4.1 Get some data

In my project, I am going to use the real time data set. It only got two dimensions, and the reason why I have chosen this is so that I can provide plots of the data what PCA analysis is doing at each step.

3.4.2 Subtract the mean

For PCA to work properly, subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have the mean of the x values of all the data points. And all the y values have the mean subtracted from them. This produces a data set whose mean is zero.

3.4.3 Calculate the covariance matrix

Since the data is two dimensional, the covariance will be 150x150. So, since the non-diagonal elements in this covariance matrix are positive, we would expect that both x and y variable increases together.

3.4.4 Calculate the Eigen vectors and Eigen values of the covariance matrix

Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues use for this matrix. These are rather important, as they tell us useful information about our data. It is important to notice eigenvectors are unit vectors. So by this process of taking Eigen vectors and So-variance matrix involves the transforming of data so that it is expressed in terms of lines.

3.4.5 Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into picture. If you look at eigenvectors and eigenvalues of the covariance matrix, you will notice Eigen values are quite different values. In fact, it turns out that the Eigen vector with highest eigenvalue with the highest eigenvalue are quite different values. In our example, the eigenvector with the large eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions.

In general, once eigenvectors are found from the co-variance matrix, the nest step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance. Now, if you like, you can decide to ignore the components of lesser significance. You do lose some information, but if the Eigen values are small, you don't lose much. If you leave out some components, the final data set will have less dimensions than original. To be precise if you originally have n dimensions in your data, and so you can calculate n eigenvectors and eigenvalues, and then you choose only the first p eigenvectors, then the final data set has only p dimensions.

What needs to be done now is you need to form a feature vector, which is just a fancy name for a matrix of vectors. This is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors, and forming a matrix with these eigenvectors in the columns.

Feature Vector = (eig1 eig2 eig3 …..eign)

3.4.6 Deriving the new data set

This is the final step in PCA, is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Final data = row Feature Vector X Row Data Adjust,
Where row Feature Vector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvectors at the top, and row Data Adjust is the mean-adjusted data transposed, i.e. the data items are in each column, with each row holding a separate dimension. I'm sorry if this sudden of all our data confuses you, but the equations from here on are easier if we take transpose of the feature vector and data first, rather that having little T symbol above their names from now on. Final Data is the final data set, with data items in columns, and dimensions along rows.

What will this give us? It will give us the original data solely in terms of the vectors we chose. Our original data set had two axes, x and y, so our data was in terms of them. /it is possible to express data in terms of any two axes that you like. If these axes are perpendicular, then the expression is the most efficient. This was why it was important that eigenvectors are always perpendicular to each vector. We have changed our data from being in terms of the axes x and y axes, and now they are in terms of our 2 eigenvectors. In case of when new data set has reduced dimensionality, i.e. we have left some of the eigenvectors out, the new data set is only in terms of the vectors that decided to keep.

To show this on our data, I have done the final transformation with each of the possible feature vectors. I have taken the transpose of the result in each case to bring the data back to the nice-like format. In case of keeping both eigenvectors for the transformation, we get data. The plot is basically the original data

The other transformation we can make is by only the eigenvector with the largest eigenvalue. As expected it has only single dimension. If you compare this data set with the one resulting from using both eigenvectors, you will notice that this data set is exactly the first column of the other.

3.4.6 Getting the old data back

Only if we took all the eigenvectors in our transformation will get exactly the original data back. If we have reduced the number of eigenvectors in the final transformation, then the retrieved data has lost some information.

Recall that the final transform is this:

Final Data = Row Feature Vector X Row Data Adjust,

Which can be turned around so that, to get the original data back,

Row Data Adjust = Row Feature Vector^{-1} X Final Data

However when we take all the eigenvectors in our feature vector. This is only true because the elements of the matrix are all unit eigenvectors of our data set. This makes the return trip to our data easier, because the equation becomes

Row Data Adjust = Row Feature Vector^{T} X Final Data

But to get the actual original data back, we need to add on the mean of that original data.

Row original Data = (Row Feature Vector^{T} X Final Data) + Original Mean

This formulas also applies to when you don't not have all the eigenvectors in the feature vector. So even when you leave out some eigenvectors, the above equation still makes the correct transform.

3.5 Univariate Compression using DWT

The Analysis of nonstationary signals often involves a comprise between how well transitions or discontinuities can be located, and how finely long-term behaviour can be identified. A typical example is the choice of window length in the short-time Fourier Transform. In Wavelet analysis one looks at the signal at different "scales" or "resolutions". A rough approximation of the signal might look stationary, while at a detailed level parent. This multi resolution, mulitscale view of signal has recently become popular.

The wavelet analysis is performed using a single prototype function called a wavelet, which can be thought of as a band pass filter. Fine temporal analysis is done with contracted (High-frequency) versions of the wavelet, while fine frequency analysis uses dilated (low-frequency) versions. The band pass filters have thus constant relative bandwidth or "constant-Q". The importance of constant relative bandwidth when perceptual processes like auditory system are involved has long been recognized; for example the musical scale introduced by bach is exponentially spaced and sub band coding of speech typically uses an octave-band splitting of signals.

In the third and final stage of compression, principal components obtained from second stage (if stage two is activated) or the screened raw MFL data obtained from the first stage (if the second stage was skipped) are transformed to wavelet domain to achieve further compression. The wavelet transform consists of projecting the signal onto a set of wave-like basis functions

$$\Psi s, \tau(t) \quad = \frac{1}{\sqrt{s}} \psi(t - \tau /S)$$

Generated from a small wave like function called the mother wave-like function called the ‘‘mother wavelet’’, Ψ (t), satisfying

$$\int_{-\infty}^{\infty} \Psi \text{ (t) } dt = 0$$

The quantities s and τ denote the scale and translation respectively. The transform of a signal with Ψs,τ (t) captures the details present in a signal at that scale (resolution) and location. Naturally these details are accompanied by a corresponding approximation, obtained by the transform of the signal with a scaling function φ (t) counterpart Ψ (t).

The distinguishing feature of wavelet transform is that it provides a resolution (mulitscale) representation (MR) of the signal similar to the representation of a geographical map at different scales (resolutions). A compact representation, desirable in compression applications, is achieved by requiring the basis functions to be orthonormal, which is achieved by choosing $s=2^j$ and $\tau = n2^j$, j ε Z. It can be shown that this MR is identical to the repeated filtering of signal through a set of (closed to ideal) low and high pass filtering [13]. The resulting coefficients are the approximation and detail coefficients respectively. Due to the nature of the wavelet basis, several sights attain sparse representations (small number of non-zero coefficients) in the wavelet domain.

For signals corrupted with noise, sparse representation is only achieved after thresholding of the wavelet coefficients to zero with and without shrinkage the significant ones. A three level wavelet decomposition of every data block received from the previous stage is implemented. The universal thresholding is used to identify the significant wavelet coefficients [10, 12]

$T = \sigma\sqrt{2} \text{ (ln (N))}; \sigma = median\ ||di||/0.6745$

Where N is the length of data array and σ is the standard deviation of the noise. For real time data $\sigma$is unknown, but can be estimated using the robust MAD estimator where {di} is the set of first level detail coefficients. Thresholding of coefficients can be carried out in two different ways – hard and soft thresholding.

Hard thresholding simply shrinks all the coefficients below the threshold to zero without affecting the significant ones. On the other hand, soft thresholding additionally shrinks the significant coefficients, thereby reducing the amplitude of signal. The amplitude of MFL signals plays an important role in determining the percentage wall loss estimation in regions containing anomalies [13]. Hence soft thresholding can lead to poor characterization of pipeline anomalies. Thus, hard thresholding is better suited for the purpose.

The significant wavelet coefficients along with their index values are retained at the end of this stage. If stage two was activated, the loading matrix (weighing to reconstruct the principal components) is also retained.

4. Result Analysis

The data presented in the report has been collected by an IPIG from an actual buried oil pipeline. One real time dataset is selected for evaluating the proposed three stage compression algorithm. Dataset consists of 64 X 5000 samples (i.e., 5000 samples from each 64 sensor).In figure 5.1 showing the plot of a channel of original data. In figure 5.2 showing the plot of µAD. Comparing both figures we can conclude that µAD captures the weld and metal defect portions.

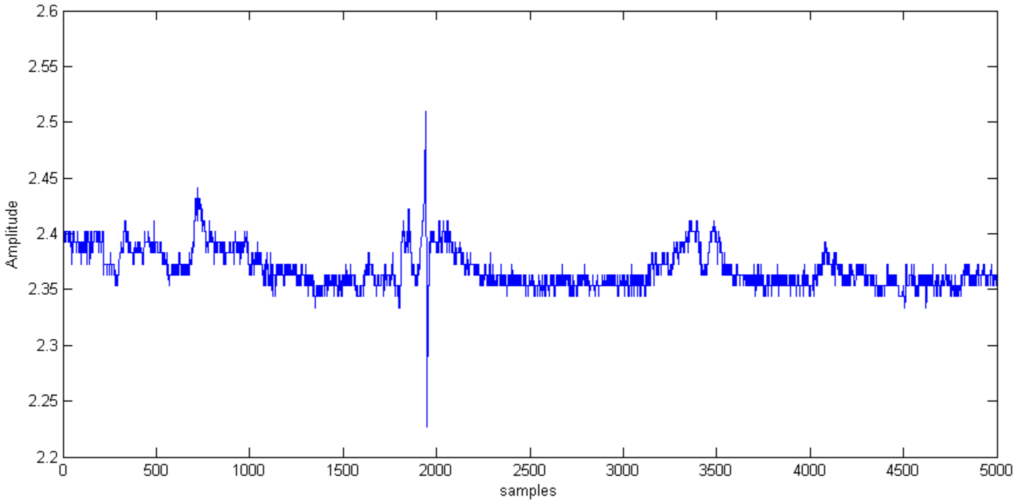


FIGURE 5.1   REAL DATA OF A CHANNEL

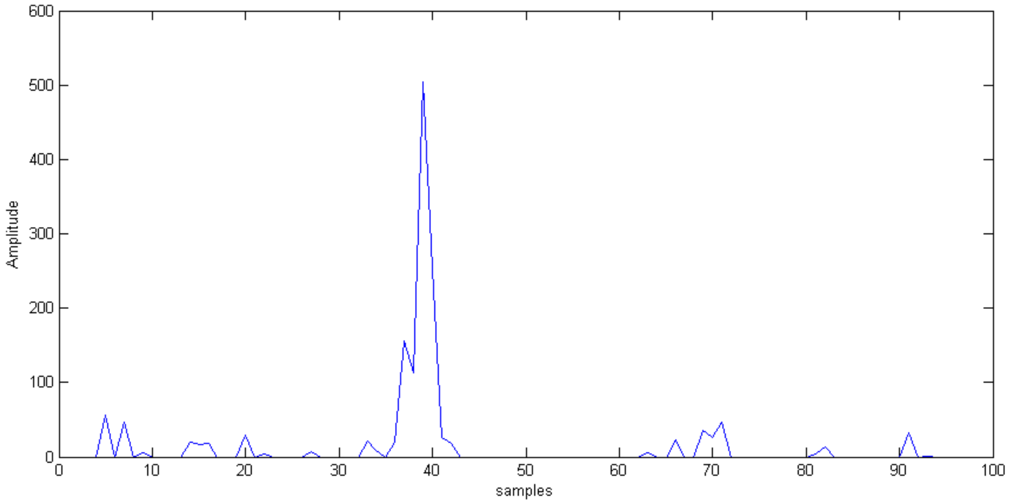


FIGURE 5.2 µAD OF CHANNEL

Reconstruction of data from the µAD is shown in fig 5.3.From the figure it is clear that we can reconstruct the data back to the original file. As we needed only weld and metal defects by using slope gradient method we can reconstruct data without losing any information. This is a lossless compression.

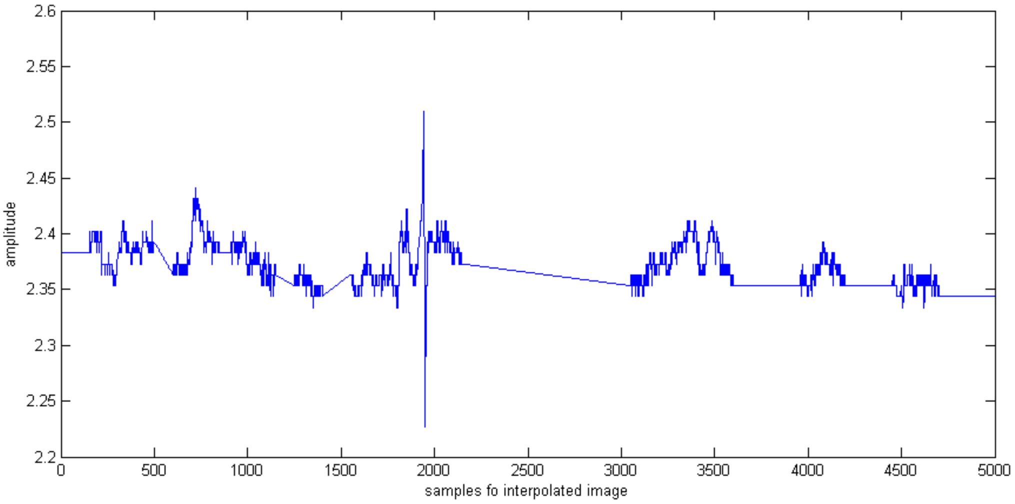


FIGURE 5.3 RECONSTRCUTED DATA BY INTERPOLATION

The data of the reconstructed page store in the compressed file is as follows

An original signal of 5000 samples is reduced into 667 samples. The compression ratio is 8. But this ratio varies with respect to defects and features. Since the features varies from page to page compression ratio won’t remain constant.

A grayscale image of the first dataset from a region near a weld is shown in fig 5.4.
A gray scale image of post screening is shown in fig 5.5.The gray blocks represents the presence of features carrying useful information about that section of the pipe. The blocks whose µAD falls below the threshold are indicated by white regions; in real time screening these data blocks are rejected. IT can be observed that the screening algorithm obtained a very few blocks with no useful information. This is due to the fact that these spurious signals have similar variability as that of MFL signals with some features. The detection algorithm is only based on variability (based on µAD) present in the signal. Thus the proposed MFL feature detection algorithm, per se, will not able to discriminate spurious signals of similar variability. Hence these signals are considered as informative blocks and processed further.

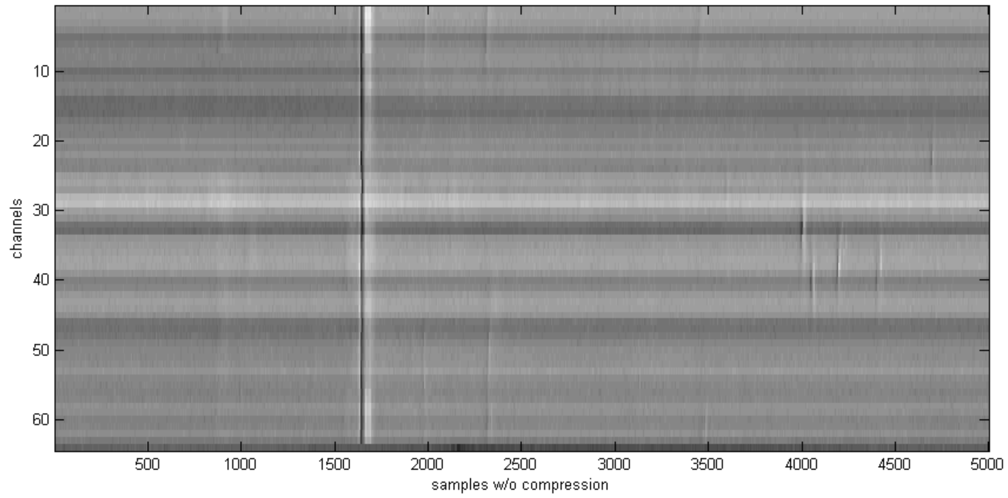


Figure 5.4 Image of the original Data

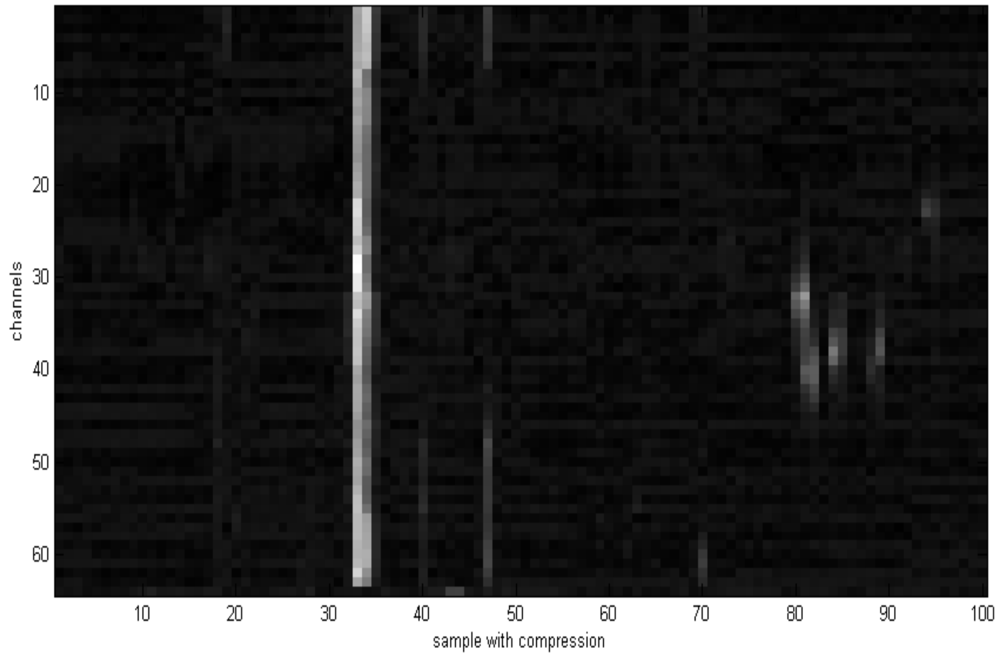


Figure 5.5 Data after screening algorithm

The data in Figure 5.5 is used normalized by adding one extra block to each side of the data block. The following data will be shown in figure 5.6. The data is showed in those values and interpolated using slope gradient formula and then the image is reconstructed back again. The compression code ratio for this image is the data stored using µAD screening algorithm 1X 16693 whereas the actual data 64X5000 so the compression ratio is 10 so it changes for every image.

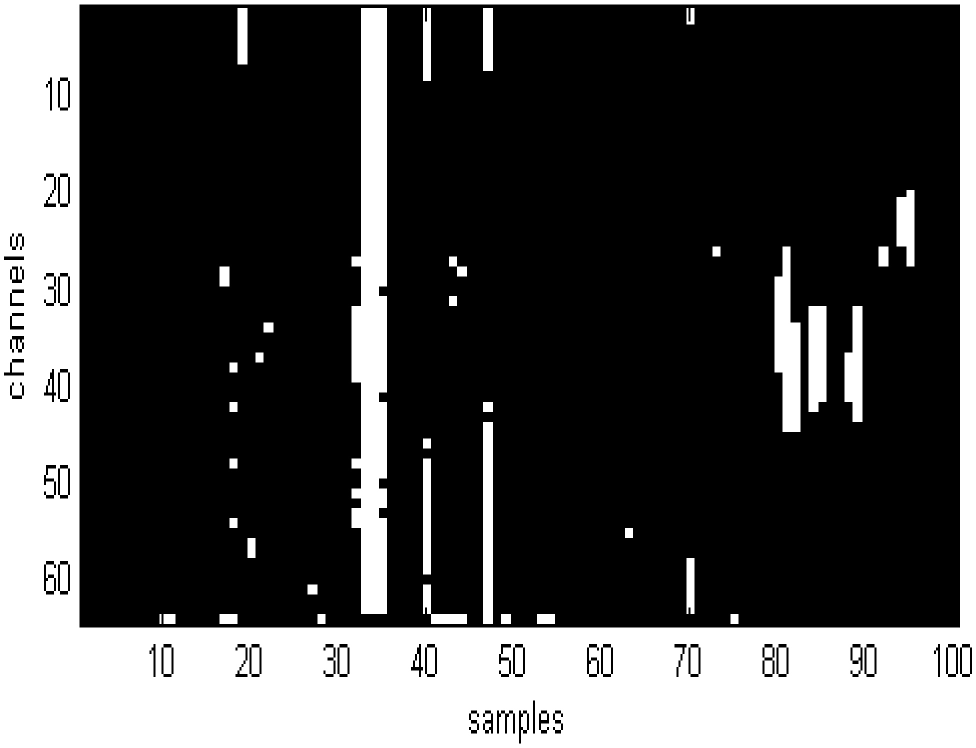


Figure 5.6 Defect features

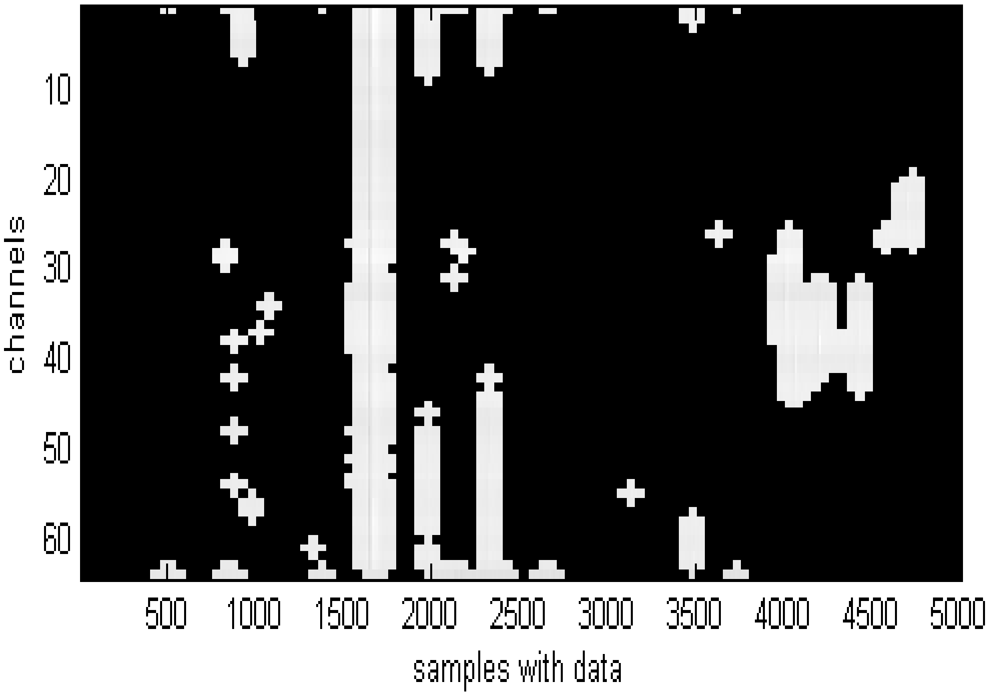


Figure 5.7 Defect Features with Data

Figure 5.19 shows the deviation between the original data and reconstructed data of weld using PCA. Since the eigenvalues are mapped linearly to original values and reducing the PCA components results in a little bit of information loss.

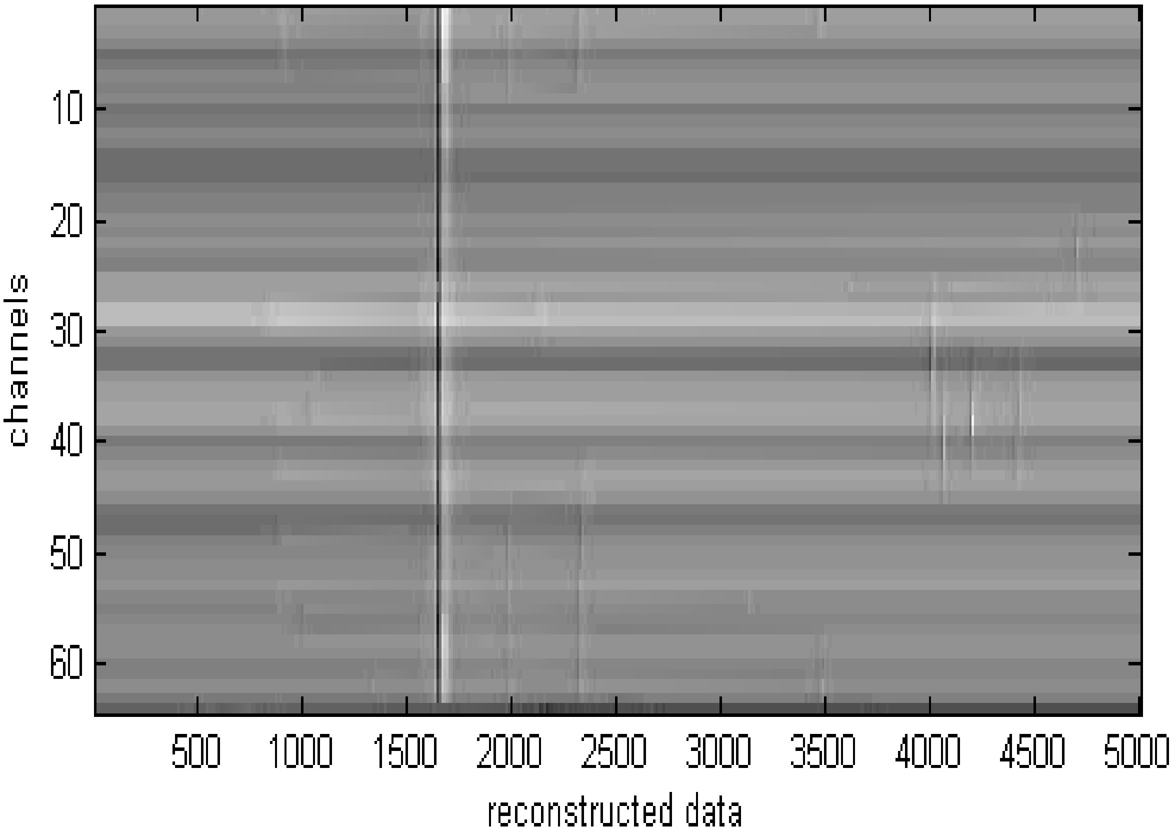


Figure 5.8 Reconstructed Data

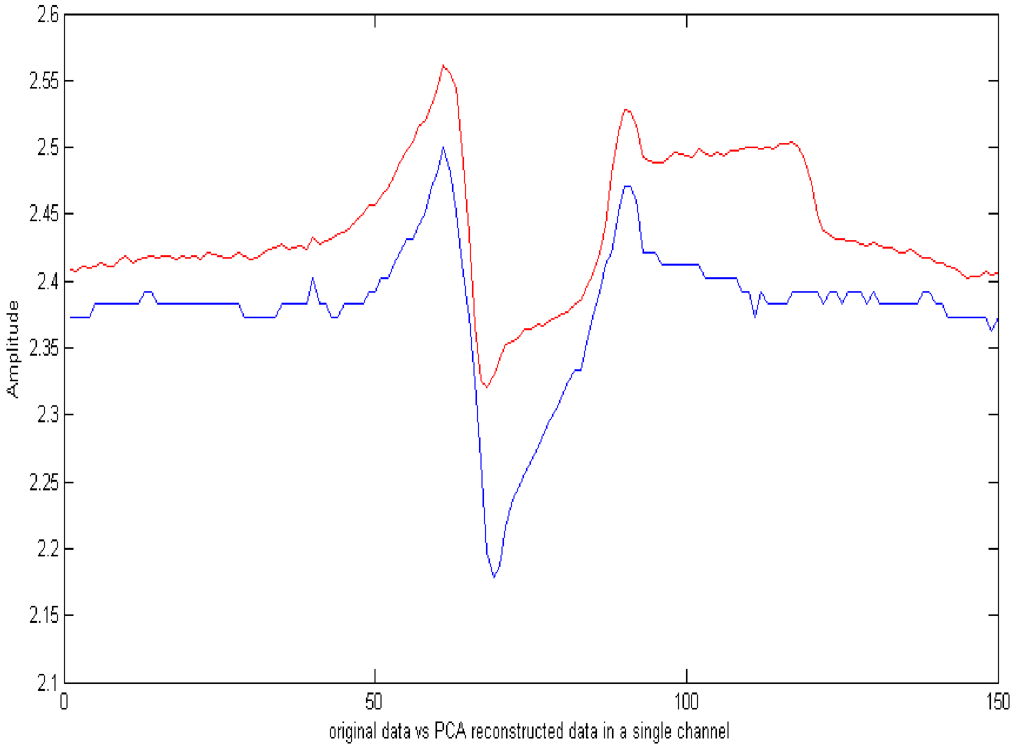


Figure 5.9 PCA Vs original Data

The Axial component of leakage flux near a defect is bell shaped. The axial component of leakage flux is in the same direction of magnetization. As the direction of magnetisation does not change in an axially magnetised pipe section the axial component of leakage flux is unipolar. The spatial distance between values of leakage flux density that are half of maximum is indicative of the length of defect. The radial component of leakage of leakage flux density, near a defect is bipolar as the direction of radial component of flux at the leading edge of defect is opposite to that at trailing edge. The distance between positive and negative peaks is the measure of length. Let us call these distances as span of axial and radial MFL signals. The peak amplitude in case of axial and directly related to the ratio of depth of defect and thickness of pipe wall. Generally an inverse relationship exists between defect length, in the range of interest, and the peak amplitude. However the influence of defect length on the peak amplitudes is required to be quantified for the magnetic circuit. Width of defect does not play a major role in sentencing a defect. But as the peak amplitudes are also directly related to width of defect does not needs to know the width for accurate estimation of depth. The secondary shape parameters, sharpness at the edges, cause significant changes in rate of change of the signal at the edges in case of axial and of the peak amplitudes in case of radial MFL signal. Sharpness at the edges in case of axial and of the peak amplitudes in case of the MFL signal. Sharpness at maximum depth has a relatively weak effect on MFL signal.

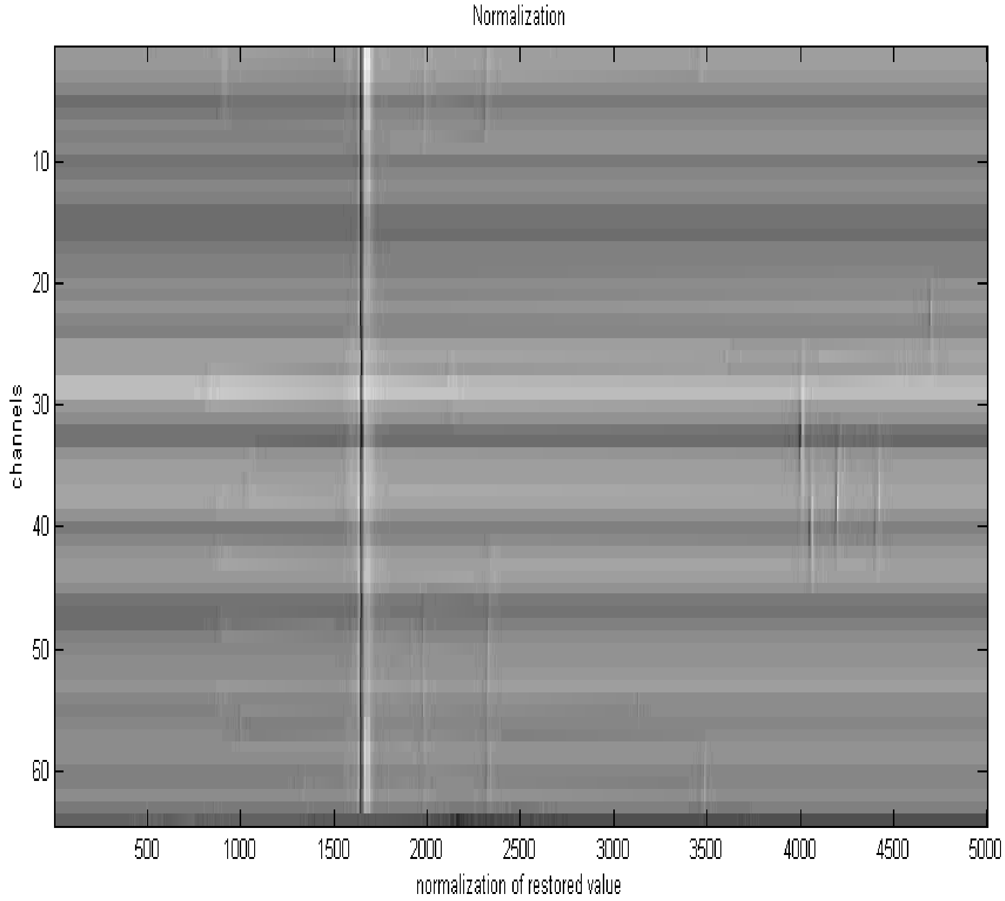


Fig 5.10 Reconstructed Images after PCA

So far the results are matching with real time data.In figure 5.11 the x-axis is PCA components and Y-axis is NMSE of whole weld data. We have PCA1, PCA3, PCA5, PCA7...PCA15 and calculated NMSE of Each channel with respect to original data. As the value of number of PCA components is rising the NMSE is slowly decreasing and at one point the NMSE is getting saturated. The point where this starts has (8% variance of whole data. So that’s why we initialized a condition for reducing how many dimensions PCA by using the variance.

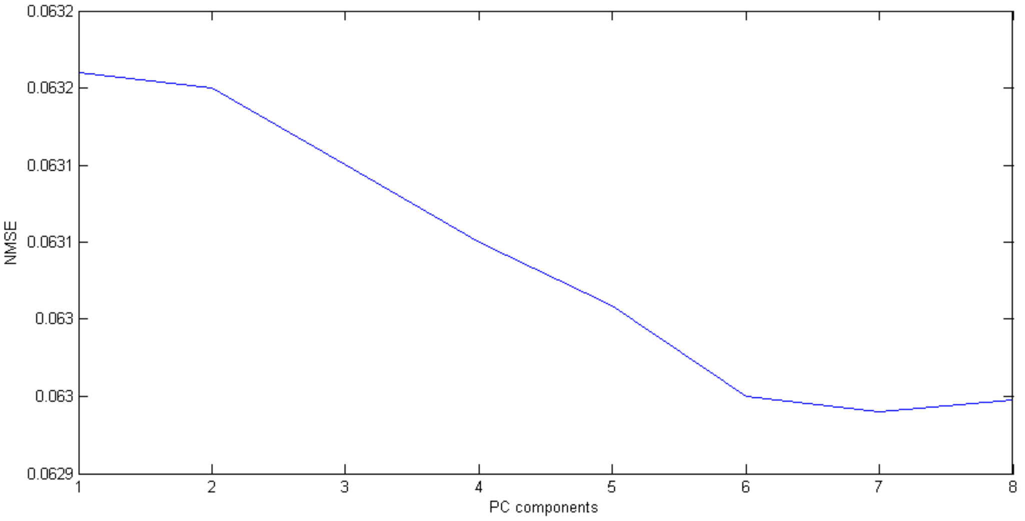


Figure 5.11 PC components Vs variance

The file for compressing the data is follows

255 page number channel block number block data... block number block data...
255 page number channel block number block data.... block number block data...
...
The procedure is similar to the above expect here the reconstruction of metal defects takes first and Weld data will be reconstructed and then the images are combined and the interpolation will take place.

Implemented the algorithm on a file which contains data of MFL signals with a memory of 23.4 MB. After implementing 1st stage on the file. The data is reduced to 2.34 MB. After implementing 2nd stage the size of file is reduced to 2.3 MB successfully. Thus the data is successfully compressed 10 times to the original file size.

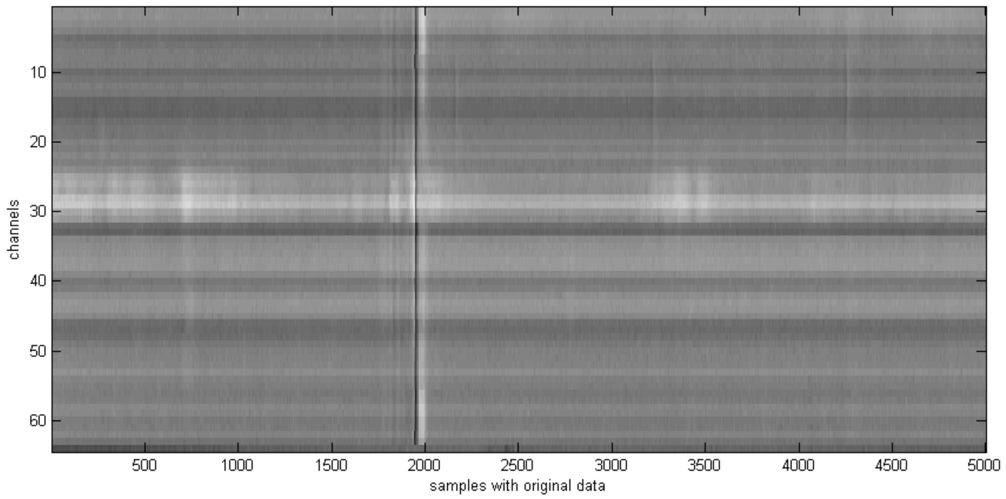


Figure 5.12 one of images in original file

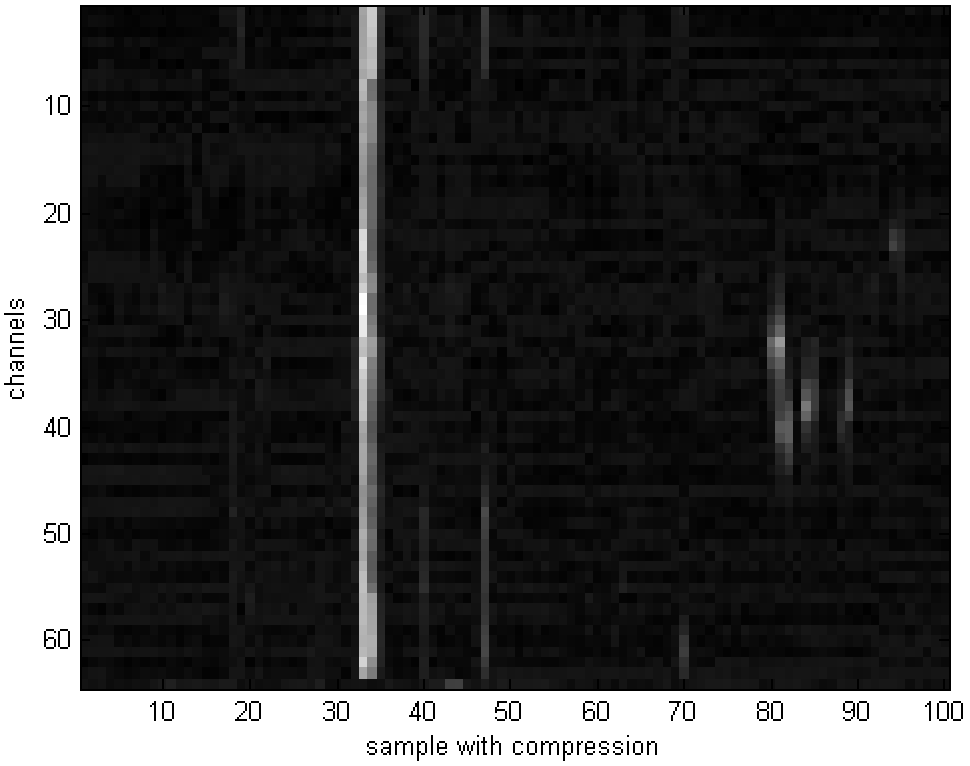


Figure 5.13 After implementation of µAD

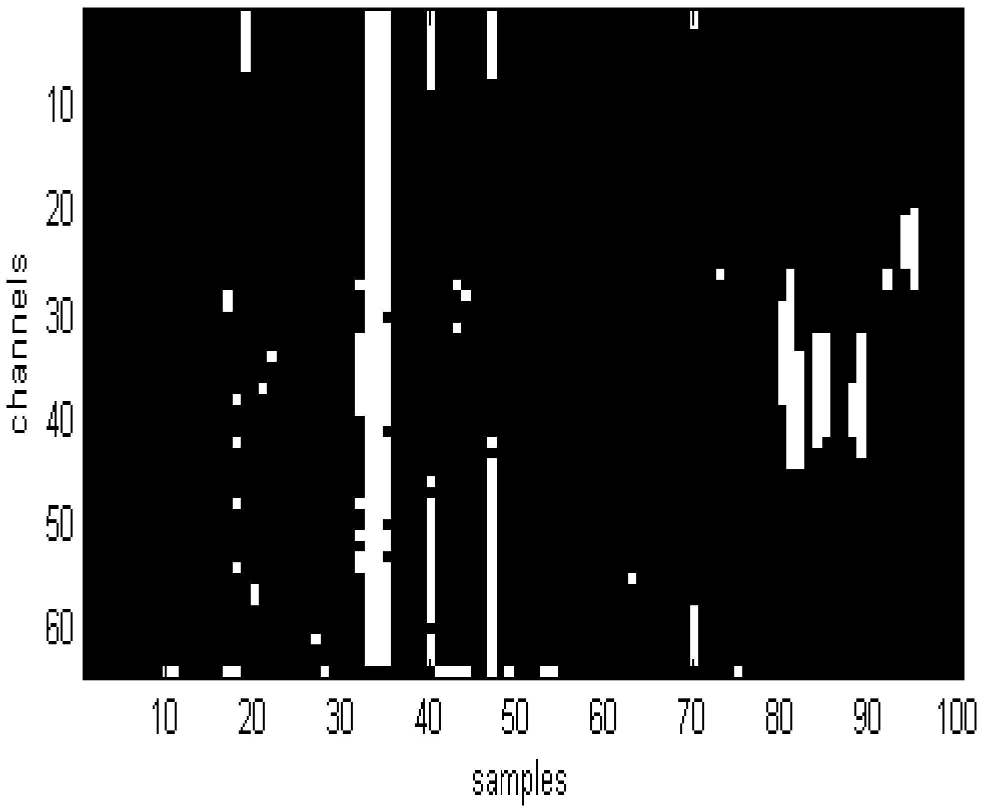


Figure 5.14 Detecting Features

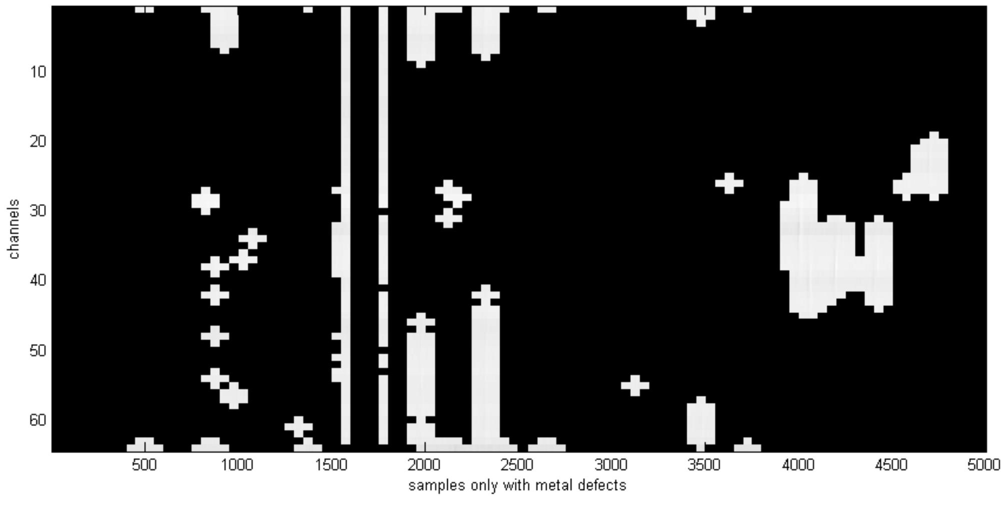


Figure 5.15 Reconstructed images only for detecting weld features

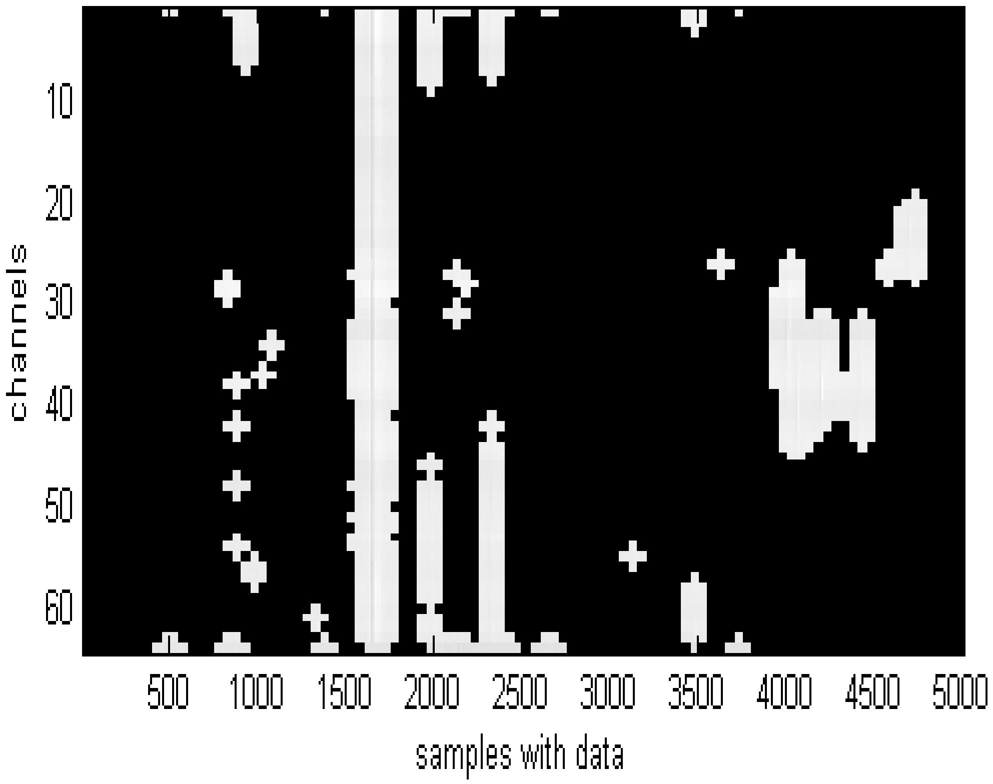


Figure 5.16 after including both weld and metal features

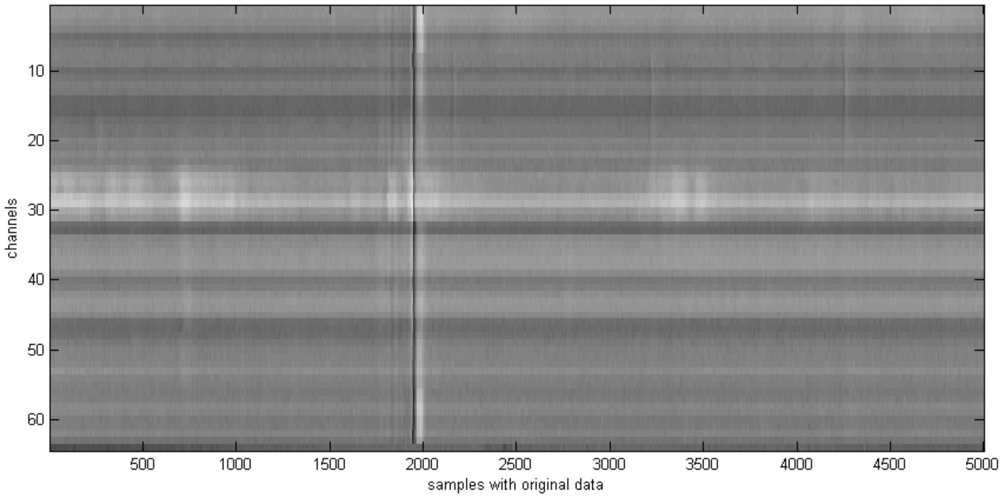


Figure 5.17 Reconstructed image after interpolation

CHAPTER 5
CONCLUSION AND FUTURE SCOPE OF WORK

5.1 Summary

Let us say there is XxYxZ Matrix. The Mean Absolute Deviation algorithm exploits the correlation among samples of single channel. The Principal component of analysis, a multivariate data compression tools is used for analysing data in lower dimension spaces. The wavelet transform signal onto a set of wave-like basis functions [13].

5.2 Conclusion

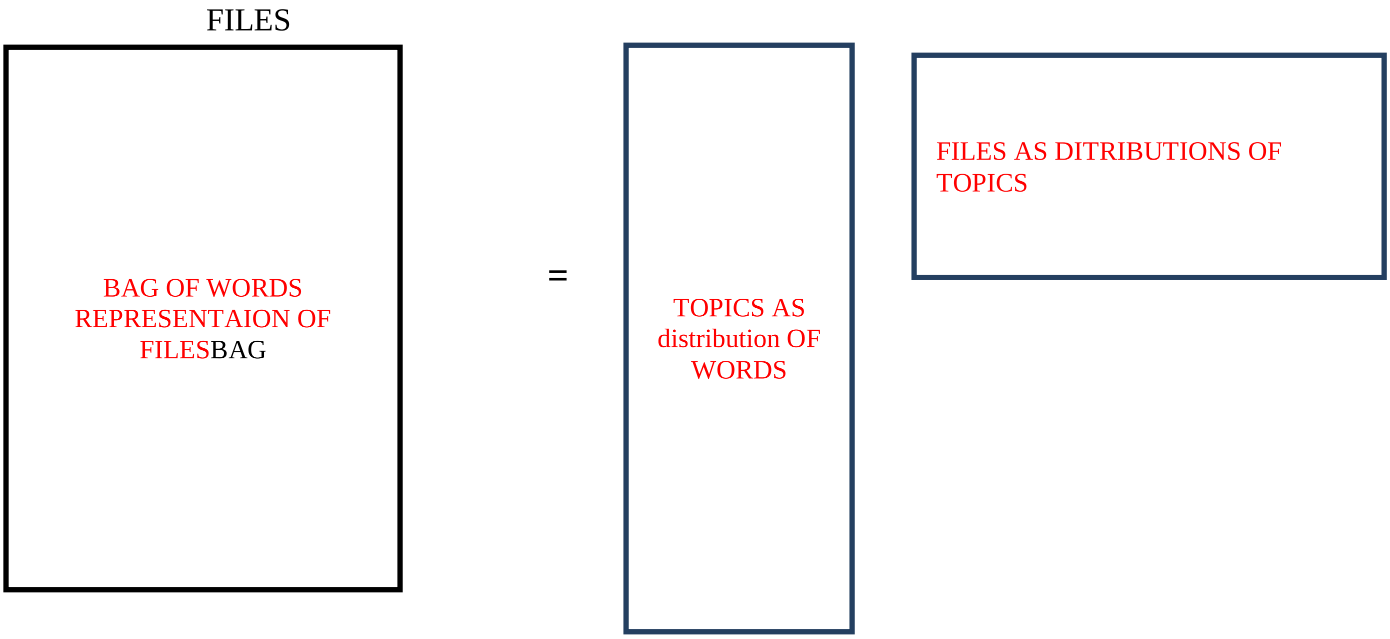
A three stage univariate-multivariate algorithm for the compression of MFL signals that can be implemented in real-time has been developed. The development stems from identifying the mathematical tools that rightly meet the requirements of the problem.

5.3 Future Scope of Work

We can improve the algorithm in two ways. One is using sparse principal component analysis. Principal component analysis (PCA) is widely used in data processing and dimensionality reduction. However, PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results. We can improve the interpretation by using sparse principal component analysis (SPCA) using the lasso (elastic net) to produce modified principal component analysis with sparse loadings. PCA can be formulated as a regression type optimization problem; sparse loadings are then obtained by imposing the lasso constraint on the regression coefficients. A new formula is to be computed for finding the total variance of modified principal component analysis.

The other is implementation of probabilistic principal component analysis. Principal component analysis (PCA) is a ubiquitous technique for data analysis and processing, but one which is not based upon a probability model. The principal axis of a set of observed data vectors may be determined through maximum-likelihood estimation of parameters in a latent variable model closely related to factor analysis. Consider the properties of the associated likelihood function, giving an EM algorithm for estimating the principal subspace iteratively.

Implementation of Discrete wavelet analysis. After the reconstruction of data, Wavelet Transform is used to further decompress the data. Implementation of Bag of words representation of documents on multiple files created after compression.


FILES
BAG OF WORDS REPRESENTAION OF FILESBAG
=
TOPICS AS distribution OF WORDS
FILES AS DITRIBUTIONS OF TOPICS

REFERENCES

[1] Srivatsava GP. Development in instrumentation and automation for NDE applications: in-House experience in the department of Atomic Energy. Insight 2003(1):73-86.

[2] Joshi A,Udpa L, Udpa S, Tamburrino A. Adaptive Wavelets for characterizing magnetic flux Leakage signals from pipeline inspection, IEEE Trans Magnet 2006:42 (10):3168-70

[3] Mandache C, Clapham L. A model for flaw detection in steel pipes by magnetic flux leakage Predictions

[4] Altschuler E., Pignotti A. Nonlinear model for flaw detection in steel pipes by magnetic flux leakage, NDT & E Int

[5] Saha S, Mukhopadhyay S.Empirical structure for characterizing metal loss defects from radial magnetic flux leakage signal, NDT & E Int

[6] Bahaguna SK, Mukhopadhyay S, Bhattacharya S, Pattil MB Development of a DSP based Data Acquisition System for IPIG project, BARC newsletter.

[7] Ziv J, Lempel A, A universal algorithm for sequential data compression. IEEE Transact. Inf Theory 1977

[8] Tipping ME, Bishop MC, probabilistic principal component analysis, J R stat soc B

[9] Tai SC, Sun CC, Yan WC. A 2D ECG compression method based on wavelet transform, IEEE Trans Biomed Engg 2005

[10] Lu Z, Kim DY, Pearlmann WA Wavelet compression of ECG by the set of partitioning in hierrarchial trees algorithm. IEEE Trans Biomed Engg 2000

[11] Jolliffe II. Principal Component Analysis, Springer; 2002.

[12] Chau FT, Liang YZ, Gao J, Shao XG. Chemometrics from basics to wavelet transform

[13] Zyoying H, Peiwen Q, Liang C. 3D FEM analysis in magnetic flux leakage method, NDT & E Int 2006.

APPENDIX A

ABBREVATION

IPIG	- instrumented pipeline inspection gauge
PCA	- Principal component analysis
EFA	- Effective Factor analysis
MFL	- Magnetic flux leakage signals
SVD	- Single value decomposition
SPCA	- Sparse Principal Component Analysis
PPCA	- Probabilistic Principal Component Analysis
DWT	- Discrete Wavelet Transform
LDA	- Linear Discriminant analysis
µAD	- Mean absolute deviation
MAD	- Median absolute deviation
FBR	- Features to Block Ratio
KLT	- Karhunen-Loeve transform
ICA	- Independent Component Analysis
NdFeB	- Neodymium Iron Boron
WHT	- Walsh Hadamard Transform
DFT	- Discrete Fourier Transform
DCT	- Discrete Fourier Transform
DDCT	- Direct Data Compression Technique
MCT	- Model Based Compression Technique
VQ	- Vector Quantization
PCT	- Parameter extraction-based compression technique
TCT	- Transform Compression Techniques
AZTEC	- Amplitude Zone Time Epoch Coding
CORTES	- Coordinate Reduction Time Encoding
EVD	- Eigen Value Decomposition.
CR	- Compression Ratio
NMSE	-Normalized Mean Square Error
FLOP	- Floating Point Operation

PROJECT DETAILS

Student Details			
Student Name	**A.Sriharsha**		
Register Number	100907551	Section / Roll No	126
Email Address	Sriharsha0806@gmail.com	Phone No (M)	9769300134
Project Details			
Project Title	**Online data compression of MFL signals for pipeline inspection**		
Project Duration	4 months	Date of reporting	07-01-2014
Expected date of completion of project	30-04-2014		
Organization Details			
Organization Name	**BHABHA ATOMIC RESEARCH CENTRE**		
Full postal address with pin code	B.A.R.C, Mumbai, India - 400 085		
Website address	www.BARC.gov.in		
Supervisor Details			
Supervisor Name	**Debmalya Mukherjee**		
Designation	Scientific officer		
Full contact address with pin code	Scientific Officer, Control Instrumentation Division, B.A.R.C. Mumbai		
Email address	debmukh@barc.gov.in	Phone No (M)	+91 22 25596120
Internal Guide Details			
Faculty Name	**Pallavi R Mane**		
Full contact address with pin code	Dept. of E&C Engg., Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address	Palvi.mane@manipal.edu		