# Design of a Suitable Framework to Recognise Inconsistencies in High Dimensional Data

## Ganji Vivekanand[1], Prof.Dr.G.Manoj Someswar[2]

1. Research Scholar, Department of CSE, VBS Purvanchal University, Jaunpur,U.P., India
2. Research Supervisor, Department of CSE, VBS Purvanchal University, Jaunpur,U.P., India

## ABSTRACT

Enormous Data investigation has pulled in exceptional premium as of late for its endeavor to remove data, learning and insight from Big Data. In industry, with the advancement of sensor innovation and Information and Communication Technologies (ICT), reams of high-dimensional, spilling, and nonlinear information are being gathered and curated to help basic leadership. The recognition of deficiencies in these information is a critical application in eMaintenance arrangements, as it can encourage upkeep basic leadership. Early disclosure of framework deficiencies may guarantee the unwavering quality and security of mechanical frameworks and lessen the danger of spontaneous breakdowns.

Complexities in the information, including high dimensionality, quick streaming information streams, and high nonlinearity, force stringent difficulties on blame identification applications. From the information demonstrating point of view, high dimensionality may cause the infamous "revile of dimensionality" and prompt decay in the exactness of blame discovery calculations. Quick streaming information streams expect calculations to give constant or close ongoing reactions upon the landing of new examples. High nonlinearity requires blame identification ways to deal with have adequately expressive power and to abstain from overfitting or underfitting issues.

Most existing flaw recognition approaches work in moderately low-dimensional spaces. Hypothetical examinations on high-dimensional blame recognition essentially concentrate on

recognizing inconsistencies on subspace projections. In any case, these models are either subjective in choosing subspaces or computationally concentrated. To meet the prerequisites of quick streaming information streams, a few techniques have been proposed to adjust existing models to an online mode to make them pertinent in stream information mining. Be that as it may, few investigations have all the while handled the difficulties related with high dimensionality and information streams. Existing nonlinear blame discovery approaches can't give palatable execution as far as smoothness, viability, heartiness and interpretability. New methodologies are expected to address this issue.

This exploration builds up an Angle-based Subspace Anomaly Detection (ABSAD) way to deal with blame discovery in high-dimensional information. The adequacy of the approach is exhibited in expository investigations and numerical delineations. In view of the sliding window system, the approach is reached out to an online mode to distinguish blames in high-dimensional information streams. Examinations on engineered datasets demonstrate the online expansion can adjust to the time-changing conduct of the checked framework and, consequently, is relevant to dynamic blame discovery. To manage very nonlinear information, the exploration proposes an Adaptive Kernel Density-based (Adaptive-KD) irregularity location approach. Numerical delineations demonstrate the approach's predominance regarding smoothness, viability and power.

*Keywords: Big Data analytics, eMaintenance, fault detection, high-dimensional data, stream data mining, nonlinear data*

## INTRODUCTION

This part depicts the exploration range; it gives the issue articulation, reason and destinations, and research inquiries of the theory, and clarifies its extension, confinements, and structure.

### Background

The across the board utilization of Information and Communication Technologies (ICT) has prompted the appearance of Big Data. In industry, remarkable rates and sizes of information are being produced from a wide exhibit of sources, including sensor-concentrated Condition Monitoring Systems (CMS), Enterprise Asset Management (EAM) frameworks, and Supervisory

Control and Data Acquisition (SCADA) frameworks. They speak to a quickly extending asset for operation and upkeep look into, particularly as specialists and experts are understanding the capability of abusing concealed an incentive from these information.
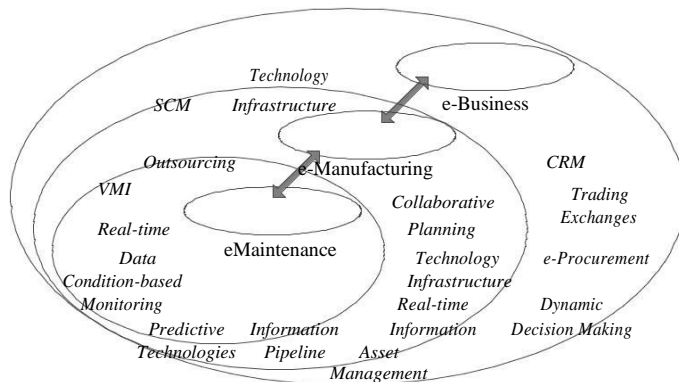


**Figure 1: Integration of eMaintenance, e-Manufacturing and e-Business systems (Koc & Lee 2003; Kajko-Mattsson et al. 2011)**

As indicated by a current McKinsey Institute report, the assembling business is one of the five noteworthy areas where Big Data examination can have transformative potential (Manyika et al. 2011). As a sub-idea of e-Manufacturing (Koc and Lee 2003), eMaintenance is likewise receiving rewards from Big Data examination (Figure 1.1 outlines the reconciliation of eMaintenance, e-Manufacturing and e-Business frameworks). One of the significant motivations behind eMaintenance is to help support basic leadership.

Through the "e" of eMaintenance, the correlated information, data, learning and insight (D/I/K/I) wind up plainly accessible and usable at the opportune place and at the ideal time to settle on the correct upkeep choices up and down the advantage life cycle (Levrat et al. 2008). This is in accordance with the motivation behind Big Data examination, which is to extricate data, information, and intelligence from Big Data.

In spite of the fact that applying Big Data examination to support basic leadership appears to be encouraging, the gathered information have a tendency to be high-dimensional, quick streaming, unstructured, heterogeneous and mind boggling (as will be point by point in Chapter 2) (Zhang and Karim 2014), in this manner posturing huge difficulties to existing information handling and investigation methods. New types of strategies and innovations are required to break down and process these information.[1] This need has spurred the improvement of Big Data examination in this postulation. To refer to (Jagadish et al. 2014): "While the potential advantages of Big Data are genuine and noteworthy, and some underlying triumphs

have just been accomplished, there stay numerous specialized difficulties that must be routed to completely understand this potential."

## Evolvement of maintenance strategy

The developing information storm has encouraged "the fourth worldview" in logical research, to be specific Data-Intensive Scientific Discovery (DISD) (Bell et al. 2009; Chen and Zhang 2014). The move from observational science (i.e., portraying regular marvels with experimental proof), hypothetical science (i.e., demonstrating of reality in view of first standards), computational science (i.e., reproducing complex wonders utilizing PCs) to DISD has been seen in different logical orders. In upkeep inquire about, a comparable progress can be found in support methodologies, as appeared in Figure 1.2.

Essentially expressed, upkeep examine has developed with the progressive substitution of Corrective Maintenance (CM) with Preventive Maintenance (PM) (Ahmad and Kamaruddin 2012). The most seasoned CM hones take after a "fall flat and fix" theory. This responsive methodology may bring about unscheduled shutdowns and prompt critical monetary misfortune or extreme dangers in wellbeing and ecological perspectives. The dread of shutdowns and their results roused organizations to perform upkeep and repair before resource disappointment, i.e., to embrace a PM system.[2] PM proposes support activities either in light of a foreordained timetable (e.g., date-book time or the utilization time of gear) or the wellbeing state of the hardware. The previous is called Predetermined Maintenance, or Time-Based Maintenance (TBM), and the last is Condition-Based Maintenance (CBM) (Ahmad and Kamaruddin 2012). In the beginning periods of PM improvement, support exercises were commonly performed at settled time interims. The PM interims depended on the information of experienced professionals and designers; the two noteworthy confinements of the approach were wastefulness and subjectivity. Another method for deciding PM interims is by following the Original Equipment Manufacturer (OEM) proposals. OEM suggestions depend on research facility analyses and dependability hypothesis, for example, Highly Accelerated Life Testing (HALT). With the entry of cutting edge figuring strategies, computational recreations of complex frameworks have likewise been utilized to suggest PM interims (Percy and Kobbacy 2000).
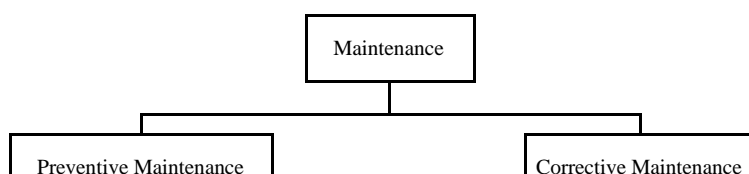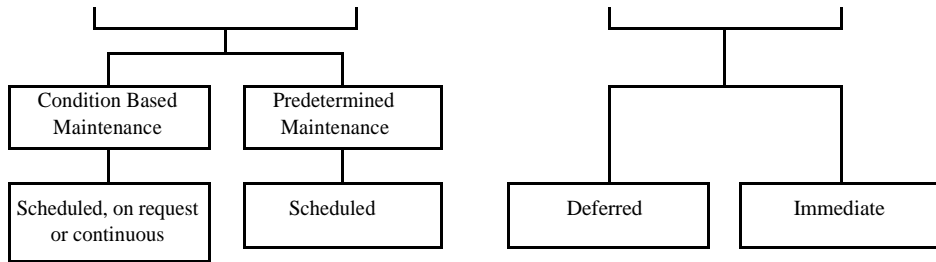
**Figure 2: Maintenance strategies** (CEN 2010)

Albeit surprising disappointments can be incredibly decreased with a foreordained support system, there are two noteworthy issues. To start with, it has a tendency to keep up gear exorbitantly,[3] causing high upkeep costs (Peng et al. 2010). It appears to be dumbfounding, however unnecessary upkeep may not really enhance the steadfastness of hardware; rather, it could even prompt more disappointments. Studies have demonstrated that 50 to 70 percent of all gear flops rashly after upkeep is completed (Karim et al. 2009). Second, it expect the disappointment conduct or trademark is unsurprising. As such, it presumes that gear disintegrates deterministically following a very much characterized grouping. Shockingly, the presumption isn't reflected actually where disappointment conduct is typically an element of hardware maturing, natural impacts, process floating, complex communications amongst segments and frameworks, and numerous different elements (Kothamasu et al. 2009). A few autonomous examinations crosswise over ventures have additionally shown that 80 to 85 percent of gear disappointments are caused by the impacts of arbitrary occasions (Amari et al. 2006).

When all is said in done, CM methodology is inclined to "deficient upkeep", while foreordained support tends towards "inordinate upkeep". To take care of the issue, a CBM procedure, or prescient upkeep, was proposed. CBM predicts future disappointments in light of the wellbeing state of hardware and starts when support errands are required. The essential distinction between foreordained upkeep and CBM is that the support exercises of the last are resolved adaptively in light of condition information. To catch the powerfully changing state of hardware, immense measures of information should be measured and gathered through condition checking, in-situ review or testing.[4] At that point, different information examination methods (e.g., machine learning, information mining, and so on.) can be connected to evaluate the wellbeing state of the gear, accordingly encouraging upkeep basic leadership.
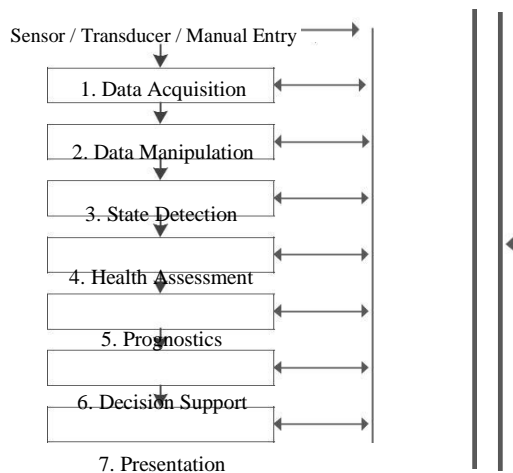
The advancement of support system speaks to a significant move from reactivity to pro-activity. It reflects the previously mentioned progress towards the DISD worldview in logical research. It

---

was empowered by hypothetical advances in upkeep administration and improvements in e-innovations. The idea eMaintenance utilizes e-innovations to help a move from "come up short and fix" support practices to "counteract and anticipate" ones (Iung and Marquez 2006). At the end of the day, it speaks to a change from the present Mean Time Between Failure (MTBF) practices to Mean Time Between Degradation (MTBD) hones (Iung and Marquez 2006).

## Condition-based support

By and by, chronicled disappointment information are regularly used to evaluate the disappointment circulation of a thing utilizing factual strategies and after that to foresee future disappointments with a specific certainty level. By and large, this works just when the concerned thing is worked in a moderately stationary condition and no unexpected changes are probably going to happen. Given the multifaceted nature of present day frameworks, different disappointment systems may associate with each other in an exceptionally complex way; ecological vulnerabilities may likewise greatly affect the event of disappointments. This requires foreseeing future disappointments of a thing in light of information which can mirror its genuine condition. It has been evaluated that 99 percent of hardware disappointments are gone before by some breakdown signs or signs (Bloch and Geitner 2012). This gives us the chance to lead CBM in view of condition estimations of the thing being referred to. The worldwide target of upkeep is to keep up a benefit's condition and expected administrations up and down its life cycle (Levrat et al. 2008). In spite of the fact that some additional expenses (e.g., sensor establishment, hand-held estimation gadget obtainment)
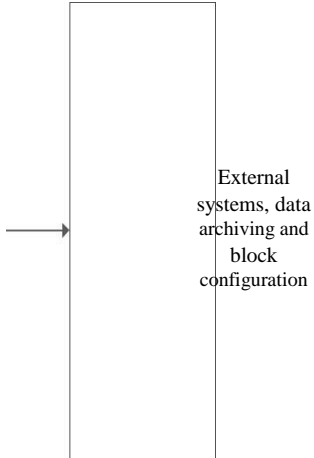


**Figure 3: OSA-CBM architecture**

External
systems, data
archiving and
block
configuration

A study uncovered that a speculation of $10,000 to $20,000 in CBM usage could prompt yearly investment funds of $500,000 (Rao 1996).[5] It has likewise been asserted that the best way to limit both support and repair costs and the likelihood of dis appointment event is to perform online framework wellbeing checking and progressing expectations of future disappointments (Kothamasu et al. 2009).

Formally, CBM is a sort of preventive support which incorporates a mix of condition observing or potentially review and additionally testing, examination and the resulting upkeep activities (CEN 2010). Condition observing, investigation, and testing are the principle part of a CBM execution. They can be led consistently, occasionally or on ask for, contingent upon the criticality of the checked thing. The resulting examination evaluates the wellbeing condition and predicts the Remaining Useful Life (RUL) of the thing; this constitutes the center of a CBM plot. The last advance – deciding support activities – includes an upkeep basic leadership process which considers support assets,[6] operational settings, and contributions from different frameworks. An Open System Architecture for Condition-Based Maintenance (OSA-CBM) has been created as per the practical determinations of ISO-13374 on condition observing and diagnostics of hardware (Swearingen et al. 2007). OSA-CBM gives a model system to CBM usage; the objective was to build up an engineering and information trade traditions that empower the interoperability of CBM parts. It is viewed as a standout amongst the most imperative models of eMaintenance frameworks (Holmberg et al. 2010). As appeared in Figure 1.3, the OSA-CBM engineering comprises of seven layers: information procurement, information control, state location, wellbeing evaluation, prognostics, choice help and introduction.[7]

x        Layer 1 – information obtaining: crude information can be aligned digitized information (e.g., sensor estimations), pictures taken from a warm imager, sound clasps taken from acoustic sensors, manual sections (e.g., writings of regular dialect) wrote by an assessor, et cetera. These information may start from various frameworks and their testing rate may contrast contingent upon the criticality of the observed thing. Step by step instructions to incorporate these information sources and lead information combination is a noteworthy test in the eMaintenance area.

x        Layer 2 – information control: this progression relates to the information planning stage in an ordinary information mining process. It covers all exercises expected to develop the last dataset for examination from the underlying crude information. Methods, for example, information purifying, information attribution, include choice, highlight extraction, and institutionalization can be connected to process the crude information to yield proper information
for promote examination. This progression is exceedingly reliant on the nature of the crude information and should be tended to contrastingly in different applications.

x        Layer 3 – state location: this progression is otherwise called blame discovery. In blame identification, information are gotten from the past advance, and their esteems are contrasted and

expected esteems or control confines; an alarm is activated if these cutoff points are surpassed. The objective of this progression can be improved to a parallel order issue, i.e., to characterize whether the thing is functioning admirably or something has turned out badly. Since the condition factors of the checked thing are subject to the operational

setting, typical conduct of the thing in one setting might be anomalous in different settings, and the other way around. In this manner, blame discovery techniques ought to know about changes in operational setting and be versatile to new operational situations.

x        Layer 4 – wellbeing appraisal: this progression concentrates on deciding whether the strength of an observed thing is debased. In the event that the wellbeing is debased,[8] an analysis on the broken condition with a related certainty level is required. Solidly, wellbeing appraisal comprises of moves made for blame acknowledgment, blame restriction, and recognizable proof of causes. The analysis method ought to have the capacity to recognize "what turned out badly" (kind, circumstance and degree of the blame) as a further examination of the way that "something turned out badly" inferred at the past advance. A wellbeing evaluation ought to likewise consider slants in wellbeing history, operational setting and support history.

x        Layer 5 – prognostics: this progression extends the conditions of the checked thing into the future utilizing a mix of prognostic models and future operational utilization models. At the end of the day, it appraises the RUL of the thing considering the future operational usage design and different elements that could influence the RUL. A certainty level of the appraisal ought to likewise be given to speak to the vulnerability in the RUL gauges.[9]

x        Layer 6 – choice help: this progression produces prescribed activities in light of the expectations without bounds conditions of the thing, present and future mission profiles, abnormal state unit targets and asset limitations. The suggested activities might be operational or support related. The previous are regularly direct, for example, notice of alarms and the consequent working techniques. On account of the last mentioned, support advisories should be nitty sufficiently gritty to plan upkeep exercises ahead of time, for example, the measure of required support work force, save parts, instruments and outside administrations.

x        Layer 7 – introduction: this progression gives an intelligent human machine interface that encourages examination by qualified staff. All the related information, data and results got in past advances ought to be associated through the system and outwardly displayed in this layer. At times, experts may require the capacity to bore down from these outcomes to get further bits of knowledge. The OSA-CBM design gives a comprehensive perspective of CBM. Each layer requires extraordinary treatment, and diverse strategies have been created particularly for each layer. Regularly, the errands characterized in these layers ought to be successively and totally completed to consequently plan condition-based upkeep undertakings. Be that as it may, sometimes, in light of an absence of information in some particular layers, the progression of this successively connected chain isn't ensured. For instance, if there are no proper prognostic models, the forecast undertaking can't be naturally performed. Under such conditions, master information[10]

and experience can simply be utilized to finish the succeeding methods. The first strategies can at present be educational and give a solid real premise to human judgments (Vaidya and Rausand 2011). In this illustration, the methodology from layer 1 to layer 4 shape the blame recognition and conclusion (FDD) application. Errands from layer 1 to layer 3 contain the blame location (FD) application, the principle look into range in this postulation.

## Problem statement

Blame recognition means to distinguish blemished states and conditions inside mechanical frameworks, subsystems and segments. As noted in the past area, the contributions of blame recognition applications are estimations mirroring the wellbeing condition of the checked thing. Since mechanical frameworks are progressively furnished with generous quantities of sensors (thermometers, vibroscopes, relocation meters, stream meters, and so on.), the state estimations have a tendency to be high-dimensional. Ordinarily, these high-dimensional estimations stream into endeavours at a rapid, in what are called quick streaming information streams. Nonlinearity is an intrinsic marvel in nature, so practically speaking, the connections between estimations can be exceedingly nonlinear. Nonlinear displaying is viewed as one of the principle challenges wherein unwavering quality meets Big Data (Göb 2013).
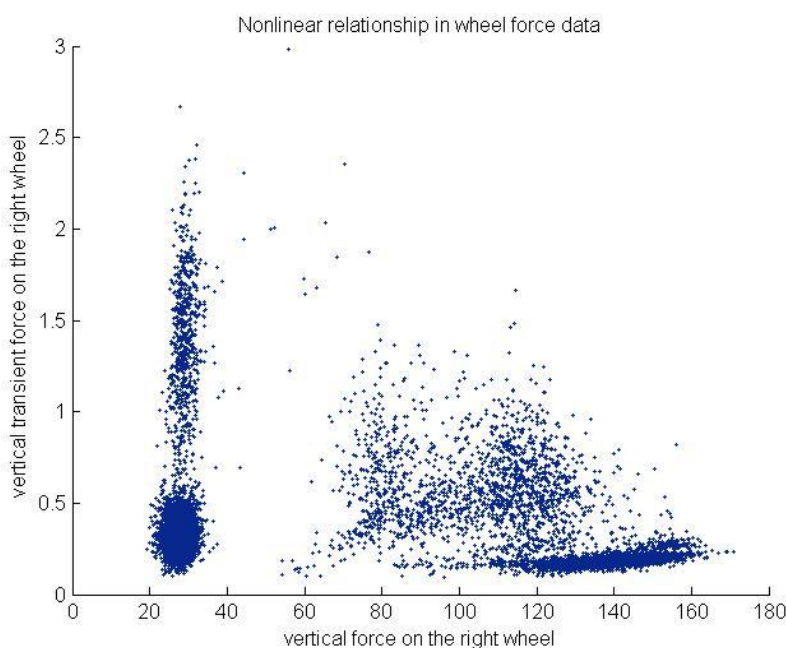


**Figure 4: High nonlinearity in a real-world dataset**

The accompanying case delineates high-dimensional information streams. In a Swedish hydropower plant, 128 simple transducers and 64 advanced transducers are sent on a hydro-generator unit. Diverse signs are caught from scattered parts of the unit, for example, rotor rotational speed, shaft control bearing temperature, water driven oil level et cetera. At present, the examining recurrence of these estimations is set to be one example for each second, so more than 30 million tuples are amassed in one year for this single generator unit. High-dimensional information streams are additionally found in the transportation part, ranger service industry, et cetera (Bahga and Madisetti 2012; Fumeo et al. 2015). A case of high nonlinearity in reality is given in Figure 4. In the figure, the vertical power on the correct wheel of a wheel-set is plotted against its vertical transient power. The figure demonstrates various

distinctive information groups with different densities. As will be, this property in the dataset can cause the precision of many blame location strategies to break down.

High dimensionality has dependably been viewed as one of the complexities of Big Data (Chen and Zhang 2014). The supposed "revile of dimensionality" may prompt decay in the exactness of conventional blame discovery calculations (Domingos 2012). Information streams reflect two other conceivably dangerous qualities of Big Data, to be specific "high speed" and "high volume".[13] Information streams require constant or close ongoing preparing; this implies blame identification calculations ought to have low figuring many-sided quality to abstain from gathering excessively information for handling in the time measurement (Gehrke 2009). Further, information streams can advance as time advances, something known as idea floating in machine learning (Zhu et al. 2010). Placing this into the setting of blame discovery, the conduct of a checked thing can shift gradually after some time – time-differing – for some reasons, including regular change, hardware maturing, and process floating. Dynamic blame location is required to oblige the common float in a non-stationary information stream (Gao et al. 2008). Nonlinearity speaks to another part of many-sided quality in Big Data. By and by, it is frequently approximated by direct (or piecewise straight) connections between highlights; see (Alippi et al. 2014) for a case. Given the unpredictability of present day frameworks, direct estimate may effortlessly underfit the issue. Exceptional treatment is required in blame recognition applications for nonlinear frameworks.

To put it plainly, high dimensionality, quick streaming information streams and high nonlinearity force stringent difficulties on blame identification applications and numerous other Big Data investigation.[14] Advances in the demonstrating of high-dimensional information streams and nonlinear information are basic. To encourage support basic leadership, displaying for e-checking, e-conclusion and e-anticipation is viewed as a vital research heading in eMaintenance (Iung and Marquez 2006). This examination concentrates on the issues related with high dimensionality, spilling information, and high nonlinearity in blame recognition applications.[15]

## Purpose and targets

This area depicts the reason and destinations of the exploration. The fundamental reason for this exploration is to examine, investigate and create ways to deal with encourage support basic leadership through eMaintenance arrangements in light of Big Data examination. All the more particularly, the exploration destinations include:

x        a demonstrate for Big Data investigation in high-dimensional support datasets, e.g.,

which can be utilized as a part of blame discovery.

x        a display for Big Data examination in high-dimensional upkeep information streams, e.g., which can be utilized as a part of online dynamic blame location.

x        a show for Big Data investigation in nonlinear support datasets, e.g., which can be utilized as a part of blame location in nonlinear frameworks.

## Research questions

To accomplish the expressed reason and targets, the accompanying exploration questions have been detailed:

RQ 1: How can designs be removed from support Big Data with high dimensionality qualities?

RQ 2: How should high-dimensional information streams be managed in the examination of upkeep Big Data?

RQ 3: How should nonlinearity be managed in the examination of support Big Data?

## Scope and limitations

The extent of this examination is the investigation of learning based, information driven blame identification methods and the improvement of models for blame location in high-dimensional information streams and nonlinear information. In particular, it creates one model for blame identification in high-dimensional information with the point of keeping up the discovery precision. The second model is an augmentation of the first with an emphasis on high-dimensional information streams. The third tends to the troubles made by high nonlinearity in the information. The approval of these models is for the most part in view of the utilization of

manufactured datasets; take note of that the information creating components of these engineered datasets have been utilized as a part of other comparative investigations. The exploration additionally looks at the execution of the proposed models with conceivable options. The first and

the third model are confirmed freely for a situation think about utilizing genuine datasets. The constraints of this theory can be portrayed as takes after. To begin with, the errands resulting from blame recognition for example, blame determination, forecast and activity proposals, are not considered, as this requires isolate investigate. Second, the information are restricted to numerical information; i.e., all out and ordinal information are not considered. Third, the manufactured datasets used to approve the proposed models are gotten and altered from other related investigations; in this manner, they may not completely uncover the benefits and inadequacies of the models. Fourth, the contextual analyses in Paper I and Paper III are restricted to sensor estimations of one particular utilitarian sub-framework, with the essential point of checking the proposed models.

## Conclusions

Based on the results of this research, the following answers have been found for, the three research questions (RQs) given in Chapter 1.

RQ 1: How can patterns be extracted from maintenance Big Data with high dimensionality characteristics?

x   The proposed ABSAD approach can select meaningful subspaces from the original high-dimensional space. In other words, it can retain dimensions which present a large discrepancy between points and their neighbouring points.

x   The analytical study proves the metric "pair wise cosine" is a bounded metric when it is used to measure vectorial angles in high-dimensional spaces, and it becomes asymptotically stable as dimensionality increases.

x  The experiments on synthetic datasets with various dimensionality settings indicate the suggested algorithm can detect anomalies effectively and has superior accuracy when compared to the specified alternatives in high-dimensional spaces.

x  The experiment on the industrial dataset shows the applicability of the algorithm in real-world fault detection applications; in addition, its feature ordering in relevant subspaces is informative to the ensuing analysis and diagnosis of abnormality.

RQ 2: How should *high-dimensional data streams* be dealt with in the analysis of maintenance Big Data?

x  The experiments on synthetic datasets indicate the ABSAD approach has the ability to discriminate low-dimensional subspace faults from normal samples in high-dimensional spaces. Moreover, it outperforms the Local Outlier Factor (LOF) approach in the context of high-dimensional fault detection.

x  The experiments on synthetic datasets further demonstrate that the sliding window ABSAD algorithm can be adaptive to the time-varying behaviour of the monitored system and produce better accuracy than the primitive ABSAD algorithm even when the monitored system has time-varying characteristics.

x  By applying the concept of trading space for time, the sliding window ABSAD algorithm can isochronously perform online fault detection.

RQ 3: How should *nonlinearity* be dealt with in the analysis of maintenance Big Data?

x  The Adaptive-KD approach is able to recognize nonlinear structures in the data.

x  The experiments on synthetic datasets demonstrate that the proposed local outlier score is a smooth measure. Further, local outlier scores of points in cluster cores are nearly identical, and those in cluster halos are significantly larger. This indicates the locality dependent kernel width can enhance the power to discriminate in anomaly detection tasks.

x  Analytical study shows that the online extension of the proposed approach is more robust to the existence of anomalies in the training set with the data refinement step. It is also more robust to changes in parameter than is the LOF approach.

x   The interpretability of the approach is much greater than other kernel methods which implicitly conduct nonlinear transformations from the input space to a feature space.

x   The experiment on the industrial dataset shows the applicability of the algorithm in real-world applications.

## Research contributions

The main contributions of this research can be summarized as follows:

x   A novel Angle-based Subspace Anomaly Detection (ABSAD) approach to high-dimensional data has been developed. The approach can be applied in industrial fault detection in high-dimensional circumstances.

x   The ABSAD approach has been extended to an online mode based on the sliding window strategy. The extension can be applied to online fault detection in a dynamic environment.

x   A novel Adaptive Kernel Density-based (Adaptive-KD) anomaly detection approach to nonlinear data has been developed. The approach has been extended to an online mode with the purpose of detecting faults from stationary, nonlinear data streams. The approach has been found superior in terms of smoothness, effectiveness, robustness, and interpretability.

### Future research

The following are considered interesting topic for future research.

x   Given the output of the ABSAD approach in fault detection applications, methods like case-based reasoning may be adopted to conduct fault diagnosis.

x   The Adaptive-KD approach can be extended to detect faults in non-stationary data streams in a temporal context, using, for example, the sliding window strategy.

x   The approaches proposed in this research can be applied to other real-world applications to verify their merits and discover and solve any shortcomings.

## REFERENCES

1. Aggarwal, C.C. et al., 2005. On high dimensional projected clustering of data streams. *Data Mining and Knowledge Discovery*, 10(3), pp.251–273.212–3220.

2. Aggarwal, C.C. & Yu, P.S., 2001. Outlier detection for high dimensional data. *ACM SIGMOD Record*, 30(2), pp.37–46.

3. Agovic, A. et al., 2009. Anomaly detection using manifold embedding and its applications in transportation corridors. *Intelligent Data Analysis*, 13(3), pp.435–455.

4. Ahmad, R. & Kamaruddin, S., 2012. A review of condition-based maintenance decision-making. *European journal of industrial engineering*, 6(5), pp.519–541.

5. Ajami, A. & Daneshvar, M., 2012. Data driven approach for fault detection and diagnosis of turbine in thermal power plant using Independent Component Analysis (ICA). *International Journal of Electrical Power & Energy Systems*, 43(1), pp.728–735.

6. Albaghdadi, M., Briley, B. & Evens, M., 2006. Event storm detection and identification in communication systems. *Reliability Engineering & System Safety*, 91(5), pp.602–613.

7. Alippi, C., Roveri, M. & Trovò, F., 2014. A self-building and cluster-based cognitive fault diagnosis system for sensor networks. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6), pp.1021–1032.

8. Alzghoul, A. & Löfstrand, M., 2011. Increasing availability of industrial systems through data stream mining. *Computers & Industrial Engineering*, 60(2), pp.195–205.

9. Amari, S.V., McLaughlin, L. & Pham, H., 2006. Cost-effective condition-based maintenance using markov decision processes. *Annual Reliability and Maintainability Symposium, 2006 (RAMS '06).*, pp.464–469.

10. Apache, Spark. Available at: https://spark.apache.org/ [Accessed October 7, 2016].

11. Bahga, A. & Madisetti, V.K., 2012. Analyzing massive machine maintenance data in a computing cloud. *IEEE Transactions on Parallel and Distributed Systems*, 23(10), pp.1831–1843.

12. Baraldi, P., Razavi-Far, R. & Zio, E., 2011. Classifier-ensemble incremental-learning procedure for nuclear transient identification at different operational conditions. *Reliability Engineering & SystemSafety*, 96(4), pp.480–488.

13. Bell, C.G., Hey, T. & Szalay, A.S., 2009. COMPUTER SCIENCE: Beyond the Data Deluge. *Science*, 323(5919), pp.1297–1298.

14. Beyer, K. et al., 1999. When is "nearest neighbour" meaningful? In *Database Theory—ICDT'99*. Springer Berlin Heidelberg, pp. 217–235.

15. Beyer, M.A. & Laney, D., 2012. The Importance of "Big Data": A Definition. *Stamford, CT: Gartner*.