# Integration of Hidden Markov Modelling and Bayesian Network for Fault Detection and Prediction of Complex Engineered Systems

Journal Pre-proof


Integration of Hidden Markov Modelling and Bayesian Network for Fault Detection and Prediction of Complex Engineered Systems

Morteza Soleimani , Felician Campean , Daniel Neagu

PII: S0951-8320(21)00331-8
DOI: https://doi.org/10.1016/j.ress.2021.107808
Reference: RESS 107808

To appear in: *Reliability Engineering and System Safety*

Received date: 17 July 2020
Revised date: 9 May 2021
Accepted date: 22 May 2021

Please cite this article as: Morteza Soleimani , Felician Campean , Daniel Neagu , Integration of Hidden Markov Modelling and Bayesian Network for Fault Detection and Prediction of Complex Engineered Systems, *Reliability Engineering and System Safety* (2021), doi: https://doi.org/10.1016/j.ress.2021.107808

# Integration of Hidden Markov Modelling and Bayesian Network for Fault Detection and Prediction of Complex Engineered Systems

Morteza Soleimani[1*], Felician Campean[1], Daniel Neagu[1]

[1]*Faculty of Engineering & Informatics, University of Bradford, UK.*

* *Corresponding author. Email: m.soleimani@bradford.ac.uk*

## Highlights

- A new approach for fault detection and prediction of engineered systems is developed
- Integration of HMM and BN can provide adequate fault detection and prediction results
- Log-likelihood values can be effectively used for data prediction
- Dynamic behaviour of the SCR system makes its reliability analysis a challenging task
- Involvement of functional knowledge is key to increase the accuracy of the results

**Abstract**

This paper presents a methodology for fault detection, fault prediction and fault isolation based on the integration of hidden Markov modelling (HMM) and Bayesian networks (BN). This addresses the nonlinear and non-Gaussian data characteristics to support fault detection and prediction, within an explainable hybrid framework that captures causality in the complex engineered system. The proposed methodology is based on the analysis of the pattern of similarity in the log-likelihood (LL) sequences against the training data for the mixture of Gaussians HMM (MoG-HMM). The BN model identifies the root cause of detected/predicted faults, using the information propagated from the HMM model as empirical evidence. The feasibility and effectiveness of the presented approach are discussed in conjunction with the application to a real-world case study of an automotive exhaust gas Aftertreatment system. The paper details the implementation of the methodology to this case study, with data available from real-world usage of the system. The results show that the proposed methodology identifies the fault faster and attributes the fault to the correct root cause. While the proposed methodology is illustrated with an automotive case study, its applicability is much wider to the fault detection and prediction problem of any similar complex engineered system.

**Keywords**: Complex engineered system; fault detection and prediction; hidden Markov modelling; Bayesian network; mixture of Gaussians; log-likelihood; automotive Aftertreatment system.

## 1. Introduction

Engineered systems, such as industrial processes, aircraft, road vehicles, manufacturing systems, electrical and electronic equipment, are becoming more complex due to the multiple interconnected subsystems/components, with high nonlinear and stochastic dynamics, include multiple control loops, operating under varying loads and noisy environments [1, 2]. Reliability and safety analysis of such systems is also becoming more challenging with the increasing complexity.

Road vehicles, as a typical example of complex engineered systems, have seen significant transformations and modernisation over the past few decades. Vehicles have become more complex following the integration of electronics, software, and networks [3], which interact dynamically to deliver the functions of the vehicle, with the vehicles now been considered cyber-physical systems [4].

Integrated health management based on diagnostics and prognostics provide significant benefits, including increase system safety and operational reliability, with significant impact on the life-cycle costs, facilitating reduced operating cost and increased revenues [5]. Diagnostics and prognostics are the main aspects of many system's health-related programs like condition-based maintenance (CBM) [6], prognostics and health management (PHM) [7], and integrated vehicle health management (IVHM) [8, 9]. Diagnostics deals with the characterisation of faults and is based on three tasks: (i) fault detection – indicating whether something is going wrong in the monitored system; (ii) fault isolation – determines which fault candidate can explain the observed abnormal behavior; and (iii) fault identification – determines the magnitude and type of the fault when it is detected [10, 11]. On the other hand, prognostics tries to estimate how likely the failure of a system is, given the current state of the system including diagnostics [10, 12].

Data-driven and physics-based approaches are the common approaches used for diagnostics and prognostics. Data-driven approaches utilise a large amount of historical data to build a model – instead of building models based on comprehensive system physics and human expertise – in order to learn the system behaviour, without a priori knowledge about the system [13]. On the other hand, physics-based approaches are underpinned by knowledge about system degradation models and the physics-of-failure in a system [7]. These approaches are suitable where the mathematical models are available for a system, based on its physical fundamentals. Compared to data-driven approaches, physics-based modellings provide a better means to handle the bias in measured data and is able to explain the behavior of a system in a wide range of operating conditions. However, it requires a deep understanding of the failure mechanisms and associated dynamics, and can be difficult to implement with complex systems with multiple competing and inter-related damage mechanisms, varying operational conditions, and multifarious data characteristics.

The research work discussed in this paper was motivated by experience with automotive propulsion systems where high variability in usage duty cycles induce significant dynamic transient behavior, which needs to be managed in order to ensure both performance and compliance with legislation. The overall objective was to carry out a study of diagnostics and prognostics techniques with the aim of introducing a data-driven methodology for fault detection, fault prediction, and isolation for the real-world operation of complex systems in general, and automotive systems in particular.

### *1.1. Review of Related Work on Data-drive Diagnostics and Prognostics*

Given the increasing availability of large volumes of data from systems in operation, data-driven approaches are increasingly employed for diagnostics and prognostics of complex engineered systems. Despite remarkable achievements, there are still significant theoretical and especially practical challenges in dealing with complex engineered systems. Characteristics of data in complex engineered systems such as nonlinearity, dynamicity, non-stationarity, and non-Gaussianity make diagnostics and prognostics more challenging tasks [13].

Data-driven algorithms such as Machine Learning methods and Bayesian-based models have proven applicability, and there is a growing trend toward the use of Artificial Neural Network (ANN)-based algorithms for diagnostics and prognostics. Deep Learning (DL) and Transfer Learning (TL) are effective at ferreting out patterns in huge sets of data and have shown potential for diagnostics and prognostics when applied to independent and identically distributed (i.i.d.) data [14]. Despite many advantages of ANN-based algorithms, one of the radical limitations is over-fitting, i.e. they struggle with causality and face difficulty to make even simple causal inferences [14]. Basically, it is assumed that these models are able to encode the general distribution of a problem into its parameters with enough examples. However, in the real-world, distributions usually are not constant because of factors that cannot be controlled and considered in the training data. Training the model with more examples may address this problem, however, growing the environment and increasing complexity make it impossible to cover all examples. This lack of understanding causality is the main reason that the application of ANN-based algorithms (especially DL and TL) is limited to superficial uses rather than real-world applications. One of the clear examples of this issue is autonomous cars which sometimes make terrible mistakes although they have been trained for many millions kilometers. Even though some efforts have been done recently to bring causality into ML, the ideas are at the conceptual level and far from reality.

When it comes to the reliability analysis of complex engineered systems causality and interrelation between components and variables cannot be ignored. Therefore, ANN-based algorithms are not good enough to deliver what is needed for diagnostics and prognostics of complex systems with complicated functional interdependencies. Bayesian-based methods – especially hidden Markov modelling (HMM) and Bayesian network (BN) – can adequately address the causality in the form of probability for real-world applications. They have shown good potential for diagnostics and prognostics of engineered systems. This is due to their capabilities to estimate both current state and future states of complex systems and their ability to represent the complex causal-effect relationship in engineered systems [15].

Hidden Markov modelling is a stochastic method for modelling a system with unobservable states. It can be represented as the simplest dynamic Bayesian network. One of the first applications of HMM for fault detection was for antenna system [16], which was based on the combination of a pattern recognition system and a finite-state HMM. Mba et.al. [17] used a combination of stochastic resonance and HMM for diagnostic and state classification in gearbox systems. In this work, the assumption was that the model parameters for observations can be accurately determined. However, because of the complexity of practical problems, this assumption seems very strong. The mixture of Gaussian model is commonly used for hidden

Markov modelling of a system with continuous data. Tobon-Mejia et.al. [18] proposed a two-step diagnostic and prognostic method (learning step and evaluation step) based on the use of the MoG-HMM. In the offline phase, degradation features of a bearing are extracted and during the online phase, models are used for diagnostic and prognostic by estimating the bearing current health state, RUL, and the associated confidence degree.

Bayesian network is a probabilistic graphical model that describes the relations between faults and symptoms, both visually and conceptually. The application of BN / dynamic BN (DBN) in dependability analyses (e.g. reliability, availability, and maintainability) has increased remarkably during the last decades owing to its modelling capabilities and advantages over other classical methods [19, 20]. Even though Bayesian network modelling is not the solution to all problems, it is fairly versatile and applicable for a wide range of applications. Ferreira et.al. [21] presented a new framework for maintenance policy in the aeronautics industry, using the Bayesian network technique. They propose a physical model and an operational model based on a Bayesian network. The physical Bayesian network simulates the behaviour of the current physical model and permits a diagnosis of the current condition of the system. Afterward, the model can be extended to an operational model that is itself a prognostic model, by using information related to an operating plan. Jun and Kim [22] proposed a new fault analysing method based on the Bayesian network for fault identification, inference, and sensitivity analysis to overcome the limitations of other methods in handling the cause-symptom relationships in fault diagnosis.

The integration of HMM and BN (or DBN) have been recently applied for diagnostics and prognostics in a few studies, however, there are some challenges. In [23], a data-driven prognostics method was presented to assess the RUL of a physical system. The authors integrated MoG-HMMs and DBNs as a modelling tool for RUL estimation, considering that degradation in the critical components of a physical system is the only cause of failure. Their method contains two learning and exploitation phases to calculate the RUL which was applied to bearings accelerated life data. In [24], HMM and BN are integrated to introduces a novel method for fault detection and diagnosis. The authors used an algorithm to classify normal operating data into danger and safe zones by dividing the data values into three parts where the middle part represents the safe zone and both upper and lower ones are related to danger values. However, this is not realistic for all condition monitoring variables. Firstly, it completely depends on the systems architecture and functional behaviour and secondly, even for a system like Tennessee Eastman (TE) chemical process, which is mostly considered as a steady-state system in the reliability analyses, there is no lower limit for variables like vibration data as the less vibration, the better. In another work [25], Don and Khan combined HMM and BN to detect and predict the identified faults in the TE process using a similar methodology. For prediction, they compared the mean of log-likelihood (LL) to find the most approximate LL from the history of training model. This method works well for detecting and predicting faults in steady-state benchmark systems where the signatures of faults are annotated; however, it might not be applicable for dynamic systems working in real-world conditions. Rebello et.al. [26] presented a new methodology for functional reliability estimation of complex industrial systems, working in a steady-state condition. They used HMM to map the continuous data into unobservable state probabilities and DBN to find the posterior system state probability by taking into account the dependency between components within a system. They assumed that the degradation in main system components is the major cause of failure in the system. However, in practice degradation is not the only cause of failure in complex systems.

### *1.2. Research Objectives and Contributions*

The motivation for this research is fault diagnosis and prognosis of open systems with continuous dynamic processes – in particular automotive systems that are complex, multi-state, transient, and dynamic. Unlike the TE process [24-26], which is a closed-loop control system with low dynamics (considered steady-state in most studies) and known signatures of faults, an automotive propulsion system has significant dynamics and the failure mechanisms are not well cataloged in terms of signatures for normal and faulty operations. Therefore, reference HMM-BN models with proven efficiency in the context of a complex system, like the one described in [24-26], are not directly applicable.

Inspired by the research works that integrated HMM and BN for diagnostics and prognostics of identified faults in the complex TE process (in particular [24-26]), this research looks at further developing the approach such that it is applicable to a continuously evolving dynamic process.

The principle of using MoG-HMM for detecting and predicting faults is to compare and exploit pattern similarity in the data via the LL values computed through HMM training. A significant variation in the pattern of LL for a new dataset is considered as a fault. Building on the idea of Don and Khan [25] who considered the mean of three adjacent log-likelihood (LL) to find the most approximate three adjacent LL from the history, an improved prediction algorithm is proposed in this work. This introduces a prediction process that considers not only the mean and standard deviation of consecutive LL values but also takes into account their increasing/decreasing patterns, to decrease the uncertainty in the prediction. To identify the root cause of a detected/predicted fault, the information provided by HMM training and LL values is propagated as empirical evidence to update the Bayesian network, built based on data and the domain expert knowledge. The probabilities obtained from updating the Bayesian network are used to isolate the related fault which is based on the virtual evidence provided by the MoG-HMM.

Given the dynamic behaviour of some engineered systems, using a static algorithm (proposed in [24, 25]) for dividing operational data into danger and safe zones is not realistic. Therefore, knowledge about functional behaviour is utilised to specify the characteristics of variables for the normal operating condition of the system.

A case study of an automotive Aftertreatment (A/T) subsystem is considered to illustrate and validate the proposed method, in conjunction with real world operating data. The Aftertreatment system (and its subsystems) is an example of complex systems that not only have the characteristics of complex systems, but also has some complexities in terms of fault detection and fault prediction [27] that go beyond some of the references that have been applied on systems such as the TE process [24-26], rolling bearings [23], or other mechanical systems [17, 28-30]. In the A/T system a few subsystems (electromechanical components with chemical reactors, and multiple sensors and software) work with complex interactions which lead to dynamic, nonlinear and non-stationary behaviours [31]. The complexity of the case study, including dealing with data collected from real world operation, provides an opportunity to validate the proposed methodology as well as to learn about challenges with diagnostics and prognostics with open systems that need to be addressed with further work.

The remainder of this paper is structured as follows: Section 2 provides a step by step description of the proposed methodology for fault detection, fault prediction and fault isolation. Section 3 introduces the automotive exhaust Aftertreatment system as a case study and presents and discusses the obtained results.

Finally, the last section of the paper includes a discussion, conclusion of the paper, and directions for future work.

## 2. Proposed Methodology for Fault detection, Prediction, and Isolation

This section introduces the generic methodology introduced for fault detection, prediction and isolation, illustrated in figure 1. The section starts with defining the technical terms and stating the assumptions, followed by a step by step description of the methodology.

### *2.1. Definitions and Model Assumptions*

Failure is an event which is referred to a transition from correct service to incorrect service, i.e., to not implementing the system required function [27]. A fault is an abnormal condition or defect at the component or sub-system level which may lead to a failure that can last for a shorter or longer period [27]. A failure can be defined only by looking at the final outcome of a system whether it is met or not. It can also be a process failure where the core process is perceived to be below an expected standard, although the activity is completed successfully.

In this study, failure and fault are defined as their generic definitions. The objective of this work is not RUL estimation and as we are dealing with complex systems, which may have dynamic behaviours, only the terms “fault detection” and “fault prediction” are used respectively to detect and predict a forthcoming fault in a complex system.

In this work, it is assumed that the model is time-homogeneous. In other words, the parameters of the conditional probability tables (CPTs) are time-invariant and transition between states does not influence the CPTs values. Gaussian distribution is a suitable model for many cases, and it is commonly used for continuous data. The assumption is that variables of the system follow the MoG distribution. Since predicting abrupt failures in a system is almost impossible, another assumption is the existence of a period before fault occurrence that provides the specific symptoms to the fault.

### *2.2. Inputs for the Methodology*

For diagnostics and prognostics of engineered system, data (e.g. condition monitoring and operational data) plays an important role. The required data for the proposed methodology includes: (i) historical data by which the CPTs for BN are generated and the data related to normal operating conditions that is used for HMM training; (ii) real-time data for process monitoring and fault detection.

Engineering knowledge about the system under study is another input, where expert opinions are exploited to support data preparation and building a Bayesian network structure. This knowledge includes information such as how the system functions, sequence of operations, functional dependencies, critical parameters for the system, and definitions of normal and faulty operating conditions.

### *2.3. Data Preparation*

For any complex system there are many operating parameters that can be involved in the analysis. However, it is efficient to identify the critical parameters and analyse the system based on these. The parameters playing more important roles in delivering the function of the system should be identified and considered as the critical parameters for the system's functionality. Input from domain experts is commonly employed to identify critical parameters.

The next step is to specify the features of the normal operating condition for a system. The term normal operating condition does not mean that a system works ideally (with 100% performance). It is used to specify a range of the parameters within that, the system delivers its main function to a high extent. (i.e. the data may contain some faults, but the system still delivers its main functions). Based on expert opinions, the attributes of the normal operating condition can be defined for the system.

In the literature, there are different classification strategies for the stages of data processing [28, 29]. Data preparation is the process of cleaning and transforming raw data before the main process for knowledge discovery from data. It normally includes data reformatting, combining data sets, removing outliers, and filling in missing values. Therefore, the operational data should be read from the data lake and converted to a suitable format (e.g. CSV format) for further processes. The next step is collecting the data for the identified critical parameters for cleaning and resampling based on the requirements of the analysis. Then, data sets representing the normal operating condition of the system are collected and combined to build a large sample of data set for the next steps of the methodology.

## *2.4. Data Analytics*

### *2.4.1. Hidden Markov Modelling*

A hidden Markov model can be defined completely by the parameters $N, D, A, B\ and\ \pi$; for simplicity we can refer to $\lambda = \{A;\ B; \pi\}$ which is called the model parameter. Where $N$ is the number of states in a hidden Markov model, $D$ is the number of distinct observations for each state, $A$ is the state transition probability, $B$ is the observation probability, and π is the initial state probability. Figure 2 gives simple illustrations of left-to-right HMM (a) and the MoG-HMM for three time slices (b) where $S_i$ is the discrete states of the system, and $M_i$ is the mixture coefficient, $Ob$ is the observation vector, $P(Ob|S_i)$ is the probability of observation, given a state and $P(S_i|S_{i-1})$ is the probability of transition from one state to another one. A MoG-HMM can be defined by the $A$ matrix, the $B$ matrix, and the initial probability $\pi$. The observation matrix $(B)$ is modelled by a Gaussian density with a mean μ, a mixture matrix $M$, and a covariance matrix $\Sigma$ ($\Sigma$ is the $D \times D$ matrix whose $(i,j)th$ entry is $\sigma_{ij} = Cov[Ob_i, Ob_j]$) [30].

The expectation-maximisation (EM) [31] is one of the commonly used algorithms to solve HMM problems, which is an iterative algorithm based on using the maximum likelihood to estimate the parameters of the statistical model with hidden variables [31].

The extensions of HMM, (e.g. Hidden Semi Markov Model, Factorial Hidden Markov Model, Layered Hidden Markov Model) have their own advantages which make them to be a good option for some specific application domains and may have better results than traditional HMM in some cases. However, considering the nature of the problem analysed in this research, traditional HMM has been chosen over other versions because of the simplicity and computational efficiency of HMM, especially in dealing with continuous data (i.e. continuous HMM).

For offline MoG-HMM training, a large data set representing the normal operating condition prepared in the first phase is used. After updating the HMM parameters using EM algorithm, LL values are computed using the forward algorithm [32] and stored for further analysis in the data prediction stage. The log-likelihood values are the probabilities that the observations ($Ob$) are generated by the given parameters of the HMM (i.e. $P(Ob|\lambda)$). The likelihood function plays a major role in statistical inference and is used in Bayesian inference and maximum likelihood parameter estimation techniques. Log-likelihood values are the natural logarithm of the likelihood values.

The next step is detecting and predicting (i.e. early identification) a potential fault from a new data set, which are the outputs of the methodology. The same process of computing LL values is performed for a new data set. The significant variation in the mean value of newly computed LL is the symptom of an abnormality and if it lasts for a noticeable period of time, it is considered as a fault. This is the fault detection part of the methodology for the system.

To predict or identify a fault earlier in the system, firstly the data sequences should be predicted which is based on LL value prediction. A sequence of $n$ LL values (i.e. $n$ sequences of data in the time unit) is predicted using the improved form of data prediction procedure introduced in [25]. The prediction process is built upon the comparison between the history of offline computed LL (LL values related to the training data set) and newly computed LL (LL values related to the new data set) which are both computed by the same trained MoG-HMM. This comparison considers two things: (i) increasing/decreasing pattern and (ii) the mean and standard deviation of consecutive LL values. Different number of consecutive LL values can be considered, however, only three consecutive values have been taken into account in this method. For example, the increasing/decreasing pattern for three (i.e. $n = 3$) consecutive LL values can have four ($2^2 = 4$) states (i.e. $LL_1 < LL_2 < LL_3$, $LL_1 < LL_2 \geq LL_3$, $LL_1 \geq LL_2 < LL_3$, and $LL_1 \geq LL_2 \geq LL_3$). To predict LL values, the mean and standard deviation of consecutive LL values for the new data set is calculated and their increasing or decreasing pattern is recorded. Then the algorithm scans the whole history of LL values to find all three consecutive LL values with the same increasing/decreasing pattern and among those, the $n$ consecutive LL values that have the closest mean and standard deviation. Figure 3 graphically represents the algorithm for this process. The assumption is that the corresponding data values for related LL values will be similar. Therefore, the LL values and their corresponding data value is considered as the prediction. Continuing this process for $n$ times will predict the next $n$ sequences of LL values and consequently the next $n$ sequences of data. The value of $n$ can be different based on the requirements of the prediction. It means that, where a short prediction is enough, it can be for example 10 or it can be 100 for a longer prediction.

Eventually, the predicted string is assessed to check the possibility of a fault in the data. To do so, by looking at the pattern of LL, the point where the mean of the signal changes most significantly is considered as the fault time. The information from LL values is sent to the trained BN as virtual evidence for the benefit of fault isolation and identifying the root cause of the fault. To compute likelihood evidence, the value of different variables in the data are assessed in a period of time before the point where the fault is identified. This is to discover the probabilities that each variable was in its danger zone. These probabilities are fed to BN as virtual evidence.

*2.4.2. Bayesian Network Modelling and Updating*

The Bayesian network is a structured knowledge representation, in which variables are considered as nodes in a graph whose structure represents the dependencies between variables. The important aspect is to learn the dependency graph of a BN. The knowledge about the causality and effect of failure modes provides an insight into possible failures and their root caused in the system. BN can calculate the posterior probabilities based on the evidence of the variables in the network, where states for some of those variables are known. Knowledge from domain experts and statistical data are two main sources of information to build a Bayesian network [15]. To build a BN structure for complex systems, the BN structure might not completely match the structure of the systems' flow network because of the acyclic nature of BN. To figure out this issue, dummy nodes can be added to the structure.

Based on the structure of the Bayesian network, historical data is used to generate CPTs for child nodes in BN. Figure 4 illustrates a sample of the algorithm for extracting condition probabilities for a two states node from a historical data set (with length $n$) where the node has only one parent node. By normalising, the sum of conditional probabilities for the child node will be "one" for each state of a parent node. (E.g. in figure 4, $P(A = True|B = True) + P(A = False|B = True) = 1$). The historical data is also used to determine the initial probabilities of parent nodes with a similar algorithm. Figure 5 shows the related algorithm for variable $B$ with $n$ records.

There are some issues related to CPTs derivation from data like limited data and missing values and more importantly, having limited information about the ranges of data for different states of a variable (e.g. safe and danger zone). To cope with these issues and to minimise the uncertainty related to providing CPTs, expert judgment can be incorporated to improve the CPTs derivation [33]. As an example, the safe and danger zones for variables, which are the base for the presented algorithms are defined based on the expert opinions about the function of the system.

In a Bayesian network, allocating values to observed variables is called evidence. Based on these pieces of evidence, belief updating is performed which amounts to computing the probability distribution over the variables of interest. For belief updating, the posterior probabilities for all nodes can be calculated using inference algorithms, when a few or even only one node is observed in real-time. This feature is one of the main advantages of BN [26].

All a BN solution needs are graph structure and CPTs [34] which both have been provided. By feeding evidence, BN is ready to update. The likelihood information is used as virtual evidence instead of hard evidence for updating the BN. Hard (regular) evidence is an observation of a random variable, being in a particular state (or having a particular value). However, observing the value that a variable has in a specific state is not always possible, or a claimed observation cannot be completely relied on. These bring uncertainty to the evidence. Virtual evidence allows entering uncertain observations (in the form of probability values over the possible states of the observation) directly into the normally unobservable variable.

After filling CPTs and feeding likelihood evidence and updating, the posterior probabilities are analysed for alternative inferences about the faults. For fault isolation, which is another output of the methodology, the method introduced in [20] is used, in which the parent node with the maximum percentage change is considered as the root cause of the fault.

### *2.5. Outputs for the Methodology*

The output of the methodology is the results of the data analytics phase, which includes fault detection, fault prediction and root cause identification. MoG-HMM training and predictive pattern analysis of LL values provide fault detection and fault prediction results and the root cause of detected/predicted faults are identified through updating BN by likelihood evidence.

## 3. Case Study: Application of Proposed Methodology for an Automotive Aftertreatment System

An automotive exhaust Aftertreatment system is a complex system consisting of several subsystems working interdependently to meet emission legislations. The complexity and dynamic behaviour of A/T make its reliability analysis a more challenging task. The reliability challenges around this system have been previously reviewed [35] highlighting the complexities associated with diagnostics and prognostics. This provides a good justification for using this system as a case study for the proposed methodology. It should be noted in this paper, only the Selective Catalytic Reduction (SCR) subsystem has been analysed and the related results have been discussed.

### *3.1. System Description and Case Study Data*

A modern automotive Aftertreatment system for a Diesel engine typically contains four different catalysts with different functions [35]: (i) Diesel Oxidation Catalyst (DOC) which oxidise HC and CO into H2O and CO2, and NO into NO2; (ii) Diesel Particulate Filter (DPF) that filters Particulate Matter (PM); (iii) SCR eliminates NO and NO2 (NOx); and (iv) Ammonia Slip Catalyst (ASC) removes excess urea (NH3) before the gases are released into the atmosphere. However, the system's configuration can be different by combining subsystems. For example, in some new A/T systems, there is not separate ASC and its function is delivered by SCR, through optimised NH3 dosing control system. Figure 6 gives a schematic representation of the A/T system, focused on the SCR subsystem. It illustrates the important variables and parameters (specified by $V\_i$) for the performance of SCR subsystem as well as the effect of each catalyst (i.e. DOC, DPF, and SCR) on the composition.

In automotive systems, measurements from various sensors are fed to Electronic Control Units (ECUs) and these data are assessed and stored by the main ECU. Typical modern cars are equipped with tens if not hundreds of ECUs and data is collected from all ECUs across the vehicle. Data is sampled periodically as well as event based. The sampling frequency can differ from 100Hz to 0.02 Hz and data may come in varying formats (e.g. dat, mf4, TDM). For an A/T system, operational data collected from various sensors along with model-based calculated data are processed and stored by ECUs.

### *3.2. Data Preparation*

*3.2.1. Function of SCR System*

SCR system converts oxides of nitrogen (NOx) into nitrogen, water, and carbon dioxide using a control system that injects urea solution through a special catalyst into the exhaust stream of a diesel engine. Equations (1) and (2) show the chemical reactions that are happening in the SCR system. In addition to nitrogen oxides and urea (which is converted to ammonia), as inputs, there are some parameters that influence the kinetic reaction model. These parameters along with inputs of the system are explained below as critical parameters for the functionality of SCR system.

$$6\,NO + 4\,NH_3 \;= 5\,N_2 + 6\,H_2O \qquad (1)$$

$$6\,NO_2 + 8\,NH_3 \;= 7\,N_2 + 12\,H_2O \qquad (2)$$

SCR is a highly nonlinear system and it operates at completely different timescales, for instance: The process of NH3 adsorption and desorption is performed at a sub-second time scale, NH3 storage can be varied at the range of minutes, and catalyst efficiency is degraded within a few years [35]. The operation of SCR system can be mainly summarised into three steps. First, the urea solution is injected. The second step includes the adsorption and desorption of NH3 and the third is the NOx reduction on the catalyst sites.

*3.2.1.1. Critical Parameters*

NOx and urea are two main inputs for the SCR system but there are some other parameters from Aftertreatment system which strongly influence SCR functions and significant variations in those parameters can affect its performance. Some of these influencing factors can be controlled (or partially controlled) by SCR system (e.g. NH3 rate). Others are external factors that are not controllable and those are controlled by other subsystems in the Aftertreatment system or powertrain system (e.g. upstream NOx level and NO2/NOx).

In this work, the critical parameters which include the systems inputs, undesired output and other parameters influencing the performance of the SCR system are considered as average SCR catalyst temperature, NH3 load and ratio, upstream NOx level, NO2 per NOx ratio, exhaust gas mass flow rate, downstream NOx level (which are respectively identified by $V\_1$, $V\_2$, $V\_3$, $V\_4$, $V\_5$ and $V\_7$ in figure 6) and degradation factor of SCR catalyst (which is related to the age of the catalyst). These critical parameters are selected based on the domain expert knowledge about the function of the system and taking into account the variables that are related to the main functions of SCR system and the variables with more participation in the potential causes of different failure modes.

*3.2.1.2. Features of Normal Operating Condition*

When there are multiple operating modes for a complex engineered system, the relationship between different measurements varies based on the process modes [36]. Therefore, it is important to specify the features of the normal operating conditions for a system.

With a foundation of knowledge about the system and its functions, the danger and safe ranges of critical parameters have been characterised for SCR system. The term normal operating condition does not mean that a system works with 100% performance. It is used to specify the safe range of the parameters within

that, the system delivers its main function without any long-term fault. In the literature (e.g. [24]), a static method is used to divide safe and unsafe ranges for a system's variables. For example, the lower 20% values and higher 20% values are considered as a danger zone and the middle part as a safe zone.
Here, the safe and danger zones for each critical parameter influencing the performance of the system are defined by expert knowledge. In the A/T system, as an example, the chemical reaction rate in the SCR system is completely dependent on temperature and both low (< 250°C) and high (> 450°C) temperatures have remarkable negative impacts on SCR's desired functions. It should be noted the low and high temperatures vary from case to case, depending on the SCR catalyst material and the mentioned low and high temperatures are just for this case study. It is worth noting that the SCR temperature is highly dependent on DPF and diesel engine functions. Another example is the NO2/NOx parameter that is a function of the DOC subsystem. At mid-temperatures, the NOx conversion is more efficient when NO2/NOx = 0.5 and for both $\ll$0.5 and $\gg$0.5 the conversion is less efficient. For some parameters, there is no lower limit and the less value for those variables, the better. Upstream NOx level variable is one of those so using the method presented in the literature for this variable is not correct and it will affect the accuracy of the results.

### *3.2.2. Cleaning and Transforming Raw ECU Data*

ECU data which was in mf4 format is read from a data repository. The variables related to the seven identified critical parameters are resampled and outliers removed. Table 1 describes these seven critical parameters (V_1 to V_7). Some data sets were merged to generate a bigger data set in order to train the HMM. Figure 7 depicts the scaled (0 to 1) distribution of this normal operating data set (containing the seven variables) which have a mixture of Gaussians distribution. This data set is used to train offline HMM. This data preparation process is also applied for new data sets, which may contain a fault, before feeding to the data analytics phase of the methodology.

## *3.3. Implementation of Data Analytics on SCR System*

The data presented in figure 7 was fed for offline HMM training. This data is fed in the form of $Ob = [V\_1, V\_2, V\_3, V\_4, V\_5,\ V\_6,\ V\_7]^T$ as an observation vector where each element is a time series related to the corresponding variable. Using the iterative EM algorithm, HMM is trained and LL values were computed and stored for the given data set for further analyses. Figure 8 depicts the algorithm training curve which shows how log-likelihood values becoming consistent after a number of iterations in EM algorithm. This means that the HMM is trained for the given data set.

It is well-known that the EM algorithms can converge to a local optimum rather than a global optimum due to local modes or saddle-points in evidence lower bound on the log-likelihood function. However, for the purpose of our approach, this is not a matter of concern as relatively short data sets are used for fault detection purposes (if working in the probabilistic domain). Thus, even though non-linearity will be present, in a relatively short time sequence it is reasonable to expect only one mode of failure, which means for a given fault condition there will be one optimum. The local or global optimum can be a matter of concern if the system runs for a very long duration (i.e. in a big test data set) which may there are multiple faults.

The open-source Bayes Net Toolbox (provided by Murphy [37]) and HMM toolbox (presented in [24]) MATLAB codes are used for HMM training and fault prediction.

The BN structure shown in Figure 9 was built for the SCR system, based on the knowledge about causality and effect and functional dependencies within the system. Considering the hierarchical arrangement of different variables for the function of the SCR system (given that downstream NOx level is the variable of interest as an output), there is linearity between the other six variables which directly influence the downstream NOx level. Seven critical parameters ($V\_1$ to $V\_7$) are considered as nodes for the Bayesian network of the SCR system and as it is clear, the variables $V\_1$ to $V\_6$ independently affect $V\_7$. A large historical data set is used for generating CPT for the child node and initial probabilities for parent nodes in the Bayesian network using the algorithms defined respectively in figure 4 and figure 5. GeNIe software is used to construct the BN structure and belief updating.

*3.4. Results and Discussion*

Before discussing the results, it is needed to define the fault for the SCR system in this work. Here, the fault for the SCR system is defined considering Euro emission legislations for NOx level that are based on milligrams per kilometer (legislations are different for the heavy-duty vehicle, light-duty vehicle [38]) and taking into account the roughly average speed of vehicles. Based on the defined terms in the previous section, faults may lead to the failure of a system or even a long-lasting fault can be considered as a failure for the system on its own. Here, the interest is to detect and predict the faults which last for more than 30 seconds as these faults may result in failure to meet the emission targets.

To detect a fault in a new data set, it is fed to the trained HMM. Figure 10 displays the computed LL values for the new fed data set. By comparing the history of LL patterns, the significant variation in the mean value of newly computed LL is the symptom of an abnormality. As it is clear, this abnormality occurs in $3305$ time units and this is a detection of the presence of a fault in the systems.

The fault prediction is based on the comparison between the history of LL values (computed using training data set) and the new computed LL values (computed for the new data set). The prediction is based on the similarity of these two LL values. Here, to predict or identify a fault earlier, the mean and standard deviation of three consecutive LL values for the new data set is calculated and their increasing or decreasing pattern is recorded. Then the model scans the whole history of LL values to find the closest mean and standard deviation values (for three consecutive LL values) with the same increasing/decreasing pattern. The corresponding data value is considered as the prediction. Doing this process 50 times ($n = 50$) will predict the next 50 sequences of LL values and of course the corresponding data. Figure 11 compares the actual LL values for the examined data and the predicted LL values for the same data. Despite some negligible spikes, it seems evident that the actual LL values follows the predicted LL values. This signifies the sufficiency of the prediction.

In figure 12, the same fault observed in figure 10 is predicted based on the LL prediction. This figure validates the accuracy of the prediction as this fault can be detected at $3305$ time units from the actual online data set (figure 10), whereas the fault is detected at $3255$ time units based on the prediction stage (figure 12), which is a noticeable improvement. It normally takes more time for a fault to grow to a detectable level in the system and this fault cannot be detected by $3305$ time units. Therefore, this method can identify the faults earlier than they happen by using the systems operational data.

To isolate or find the root cause of the detected fault, the trained BN is updated by feeding the likelihood evidence. The values of all variables in the data are assessed in a period of time before the point where the fault is detected in order to find the probabilities of being in the danger zone for each variable. These probabilities are used as virtual evidence to update BN. After updating BN, the parent node with the maximum percentage change is considered as the root cause of the fault. Table 2 compares the prior and posterior faulty state probabilities after updating BN with the new virtual evidence. As it is obvious, variable 1 (average SCR catalyst temperature) is the cause of the fault observed in figure 10 (or predicted in figure 12).

The analysed data set (i.e. new data set) to detect and predict the fault was a data set related to the condition that temperature goes beyond the acceptable level and affects the performance of the SCR system. Therefore, the root cause is identified correctly.

To validate the obtained results and to show the efficiency of involving engineering knowledge in data analysing, the result of fault isolation is compared for two cases: with and without considering engineering knowledge in dividing safe and danger zones for variables. Table 3 compares the initial and updated probabilities by feeding virtual evidence for two cases. As discussed, $V\_1$ is the actual root cause of the fault. However, by using the static algorithm presented in the literature (e.g. in [24]) for defining safe and danger zones, the maximum percentage change is for $V\_4$ (NO2/NOx) and that is introduced as the root cause of the fault which is not correct. This shows the key impact of engaging engineering knowledge in the data analysis.

## 4. Discussion and Conclusion

The objective of this work was to introduce a methodology to detect, predict, and isolate faults in complex engineered systems with real-world operating data. To achieve this objective, a methodology based on the integration of HMM and BN was introduced, with a predictive pattern mining algorithm for detecting and predicting faults in open systems with continuous dynamic processes. The methodology was illustrated through the implementation on an automotive Aftertreatment system, which provided good preliminary validation of the methodology, proving feasibility to detect and predict unidentified faults in dynamic processes.

This work was inspired by previous results [24-26] dealing with diagnostics and prognostics of identified faults on the TE benchmark process, and demonstrated that the integration of HMM and BN for fault analysis can be extended to dynamic processes where the signatures of the faults are not known. Unlike the TE process (which is normally considered as a steady-state system since it has a very slow dynamic) automotive propulsion systems involve continuous and dynamic processes, for which failure mechanisms are not completely analysed and known, and therefore the signatures of normal and faulty operations are not always available. To address the features of dynamic processes and to decrease the prediction uncertainty in using historic data, the proposed methodology takes into account the standard deviation and increasing/decreasing patterns of data, as well as the mean values. In addition, knowledge about the functional behaviour of the system under study is utilised to specify the ranges of critical parameters for operational data of the system. The illustrative example shows that consideration of functional knowledge of the system under study is essential for the validation of the correct root cause identification for the fault.

The methodology would be further advanced through with a more formal approach to extracting knowledge from engineering analyses to support the data-driven method and to provide justification and validation to the obtained results.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### CRediT authorship contribution statement

**Morteza Soleimani:** Conceptualization, Investigation, Methodology, Data curation, Validation, Software, Visualization, Writing - original draft.

**Felician Campean:** Project administration, Conceptualization, Methodology, Supervision, Writing - review & editing.

**Daniel Neagu:** Project administration, Methodology, Supervision, Writing - review & editing

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

The authors would like to thank the anonymous reviewers whose constructive feedback helped to improve this article.

## References

1. Jiang, Y. and S. Yin, *Recursive total principle component regression based fault detection and its application to vehicular cyber-physical systems.* IEEE Transactions on Industrial Informatics, 2017. **14**(4): p. 1415-1423.
2. Bar-Yam, Y., *Dynamics of complex systems*. 2019: CRC Press.
3. Campean, F., et al. *Automotive IVHM: Towards Intelligent Personalised Systems Healthcare*. in *Proceedings of the Design Society: International Conference on Engineering Design*. 2019. Cambridge University Press.

4. Nunez, D.L. and M. Borsato, *An ontology-based model for prognostics and health management of machines.* Journal of Industrial Information Integration, 2017. **6**: p. 33-46.
5. Sun, B., et al., *Benefits and challenges of system prognostics.* IEEE Transactions on reliability, 2012. **61**(2): p. 323-335.
6. Alaswad, S. and Y. Xiang, *A review on condition-based maintenance optimization models for stochastically deteriorating system.* Reliability Engineering & System Safety, 2017. **157**: p. 54-63.
7. Lee, J., et al., *Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications.* Mechanical systems and signal processing, 2014. **42**(1-2): p. 314-334.
8. Benedettini, O., et al., *State-of-the-art in integrated vehicle health management.* Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2009. **223**(2): p. 157-170.
9. *Design & Run-Time Information Exchange for Health-Ready Components JA6268_201804*, in *HM-1 Integrated Vehicle Health Management Committee.* https://doi.org/10.4271/JA6268_201804. 2018, SAE International.
10. Sikorska, J., M. Hodkiewicz, and L. Ma, *Prognostic modelling options for remaining useful life estimation by industry.* Mechanical systems and signal processing, 2011. **25**(5): p. 1803-1836.
11. Belkacem, L., et al., *Diagnostic and prognostic of hybrid dynamic systems: Modeling and RUL evaluation for two maintenance policies.* Reliability Engineering & System Safety, 2017. **164**: p. 98-109.
12. Kan, M.S., A.C. Tan, and J. Mathew, *A review on prognostic techniques for non-stationary and non-linear rotating systems.* Mechanical Systems and Signal Processing, 2015. **62**: p. 1-20.
13. Zhong, K., M. Han, and B. Han, *Data-driven based fault prognosis for industrial systems: a concise overview.* IEEE/CAA Journal of Automatica Sinica, 2019.
14. Schölkopf, B., et al., *Toward Causal Representation Learning.* Proceedings of the IEEE, 2021.
15. Langseth, H. and L. Portinale, *Bayesian networks in reliability.* Reliability Engineering & System Safety, 2007. **92**(1): p. 92-108.
16. Smyth, P., *Hidden Markov models for fault detection in dynamic systems.* Pattern recognition, 1994. **27**(1): p. 149-164.
17. Mba, C.U., et al., *Condition monitoring and state classification of gearboxes using stochastic resonance and hidden Markov models.* Measurement, 2018. **126**: p. 76-95.
18. Tobon-Mejia, D.A., et al., *A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models.* IEEE Transactions on reliability, 2012. **61**(2): p. 491-503.
19. Weber, P., et al., *Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas.* Engineering Applications of Artificial Intelligence, 2012. **25**(4): p. 671-682.
20. Amin, M.T., S. Imtiaz, and F. Khan, *Process system fault detection and diagnosis using a hybrid technique.* Chemical Engineering Science, 2018. **189**: p. 191-211.
21. Ferreiro, S., et al., *Application of Bayesian networks in prognostics for a new Integrated Vehicle Health Management concept.* Expert Systems with Applications, 2012. **39**(7): p. 6402-6418.
22. Jun, H.-B. and D. Kim, *A Bayesian network-based approach for fault analysis.* Expert Systems with Applications, 2017. **81**: p. 332-348.
23. Medjaher, K., D.A. Tobon-Mejia, and N. Zerhouni, *Remaining useful life estimation of critical components with application to bearings.* IEEE Transactions on Reliability, 2012. **61**(2): p. 292-302.
24. Don, M.G. and F. Khan, *Dynamic process fault detection and diagnosis based on a combined approach of hidden Markov and Bayesian network model.* Chemical Engineering Science, 2019. **201**: p. 82-96.
25. Don, M.G. and F. Khan, *Process Fault Prognosis Using Hidden Markov Model-Bayesian Networks Hybrid Model.* Industrial & Engineering Chemistry Research, 2019. **58**(27): p. 12041-12053.
26. Rebello, S., H. Yu, and L. Ma, *An integrated approach for system functional reliability assessment using Dynamic Bayesian Network and Hidden Markov Model.* Reliability Engineering & System Safety, 2018. **180**: p. 124-135.

27. Avizienis, A., et al., *Basic concepts and taxonomy of dependable and secure computing.* IEEE transactions on dependable and secure computing, 2004. **1**(1): p. 11-33.
28. Jardine, A.K., D. Lin, and D. Banjevic, *A review on machinery diagnostics and prognostics implementing condition-based maintenance.* Mechanical systems and signal processing, 2006. **20**(7): p. 1483-1510.
29. Sagiroglu, S. and D. Sinanc. *Big data: A review*. in *2013 international conference on collaboration technologies and systems (CTS)*. 2013. IEEE.
30. Murphy, K.P., *Dynamic bayesian networks: representation, inference and learning*. 2002, Ph.D. Thesis, University of California.
31. Moon, T.K., *The expectation-maximization algorithm.* IEEE Signal processing magazine, 1996. **13**(6): p. 47-60.
32. Rabiner, L.R., *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 1989. **77**(2): p. 257-286.
33. Hu, J., et al., *An integrated method for safety pre-warning of complex system.* Safety science, 2010. **48**(5): p. 580-597.
34. Rohmer, J., *Uncertainties in conditional probability tables of discrete Bayesian Belief Networks: A comprehensive review.* Engineering Applications of Artificial Intelligence, 2020. **88**: p. 103384.
35. Soleimani, M., F. Campean, and D. Neagu, *Reliability Challenges for Automotive Aftertreatment Systems: a State-of-the-art Perspective.* Procedia Manufacturing, 2018. **16**: p. 75-82.
36. Ma, J. and J. Xu, *Fault prediction algorithm for multiple mode process based on reconstruction technique.* Mathematical Problems in Engineering, 2015. **2015**.
37. Murphy, K. *Hidden Markov Model (HMM) Toolbox for Matlab* 1998 Accessed Jun 30, 2020]; Available from: https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html].
38. Hooftman, N., et al., *A review of the European passenger car regulations–Real driving emissions vs local air quality.* Renewable and Sustainable Energy Reviews, 2018. **86**: p. 1-21.

**Table 1**. Critical parameters considered for the SCR system

| **Variable** | **Description** |
|---|---|
| $V\_1$ | Average SCR catalyst temperature |
| $V\_2$ | NH3 load and ratio |
| $V\_3$ | Upstream NOx level |
| $V\_4$ | NO2 per NOx ratio |
| $V\_5$ | Exhaust gas mass flow rate |
| $V\_6$ | Degradation factor of SCR catalyst |
| $V\_7$ | Downstream NOx level |

**Table 2.** The change in the probability of each node, after BN updating

| Node | Initial probability | Updated probability |
|---|---|---|
| $V\_1$ | **0.33** | **0.92** |
| $V\_2$ | 0.15 | 0.25 |
| $V\_3$ | 0.19 | 0.43 |
| $V\_4$ | 0.27 | 0.42 |
| $V\_5$ | 0.14 | 0.22 |
| $V\_6$ | 0.00 | 0.00 |
| $V\_7$ | 0.30 | 0.41 |

**Table 3.** Comparing the probabilities with and without considering system engineering knowledge

| | with considering system engineering knowledge | | without considering system engineering knowledge | |
|---|---|---|---|---|
| Node | Initial probability | Updated probability | Initial probability | Updated probability |
| $V\_1$ | **0.33** | **0.92** | 0.68 | 0.99 |
| $V\_2$ | 0.15 | 0.25 | 0.54 | 0.75 |
| $V\_3$ | 0.19 | 0.43 | 0.58 | 0.83 |
| $V\_4$ | 0.27 | 0.42 | **0.58** | **0.98** |
| $V\_5$ | 0.14 | 0.22 | 0.50 | 0.58 |
| $V\_6$ | 0.00 | 0.00 | 0.00 | 0.00 |
| $V\_7$ | 0.30 | 0.41 | 0.42 | 0.78 |

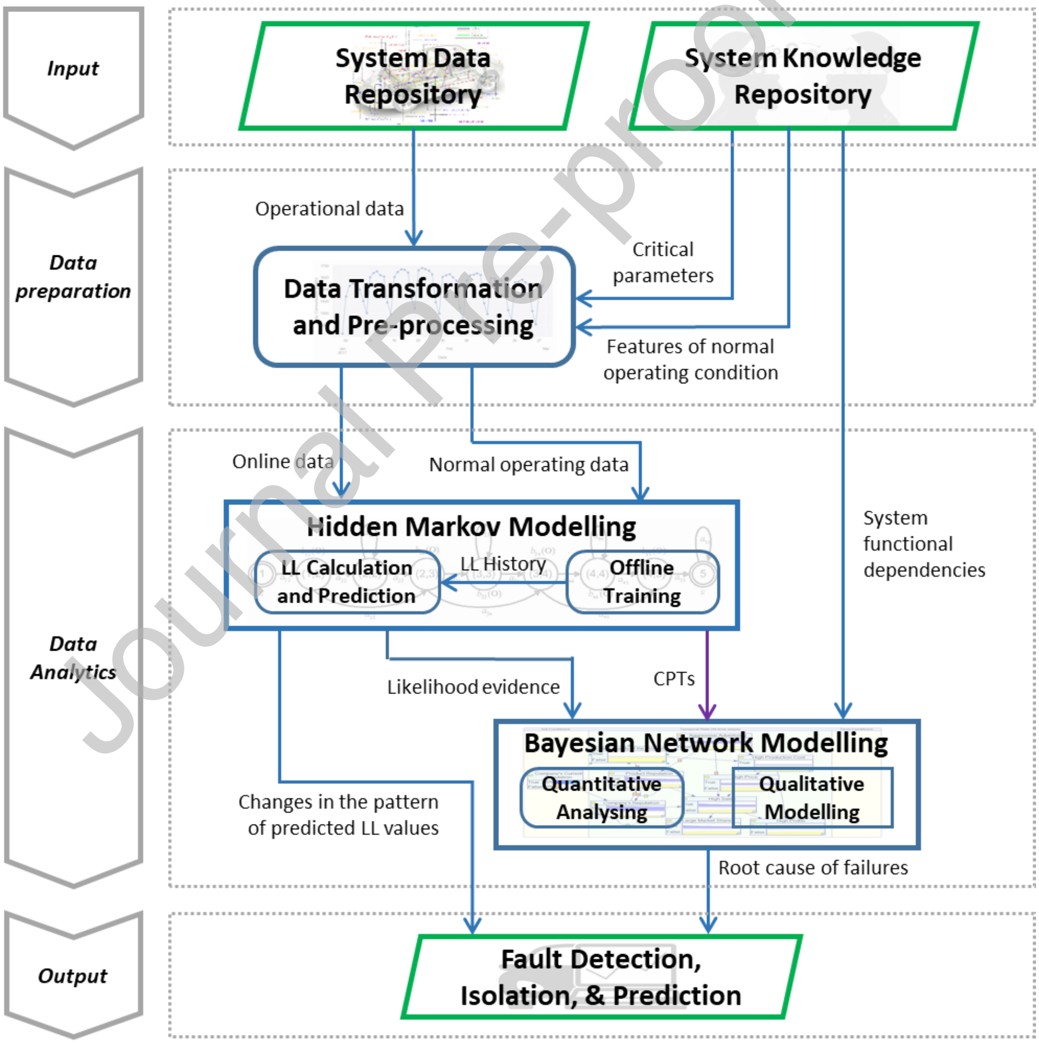
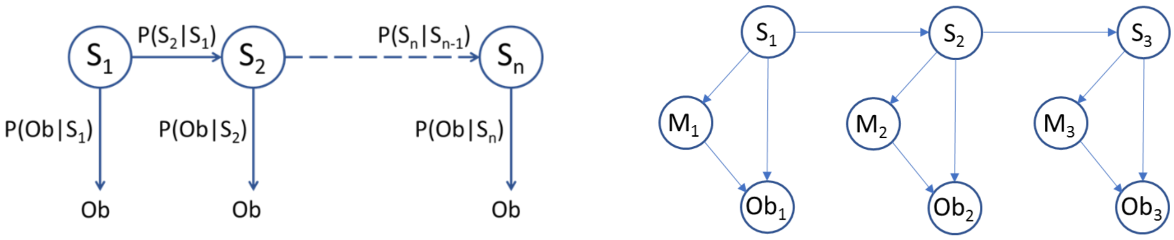



**Figure 1.** Proces




(a) (b)

**Figure 2.** Representation of left-to-right HMM (a), and a MoG-HMM for three time slices (b)

s flow diagram of the proposed methodology

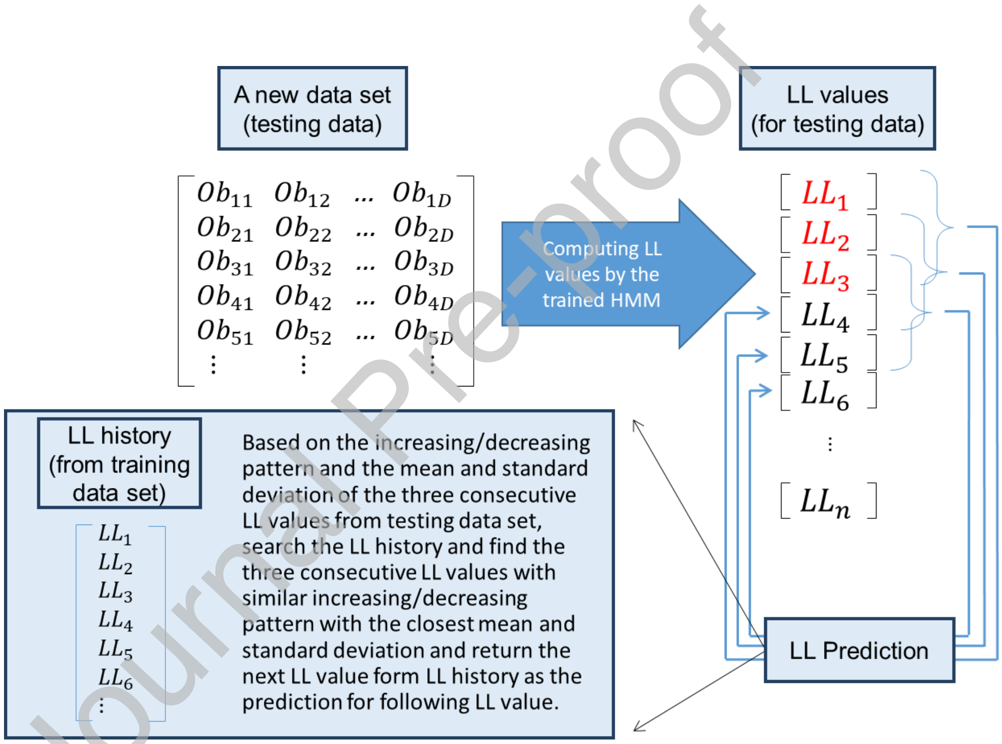
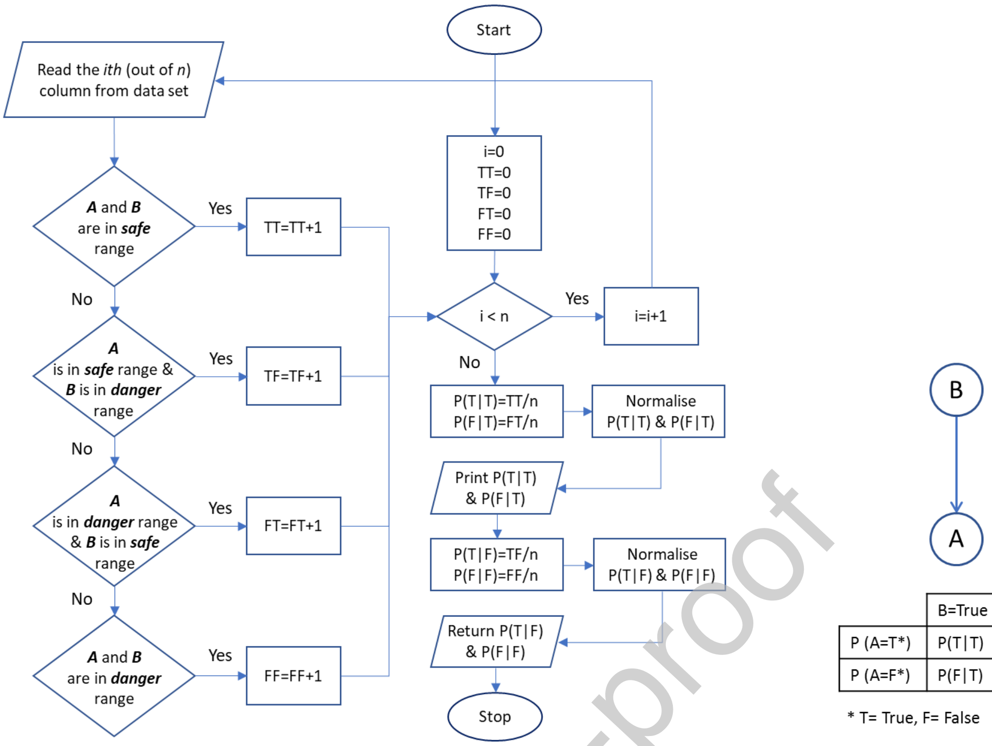



Figure 3. The graphical representation of the algorithm for LL prediction, based on three consecutive values




| | B=True | B=False |
|---|---|---|
| P (A=T*) | P(T\|T) | P(T\|F) |
| P (A=F*) | P(F\|T) | P(F\|F) |

* T= True, F= False

(a) (b)

**Figure 4.** The algorithm of extracting CPT from historical data (a) for a node *A* with one parent node, *B* (b)

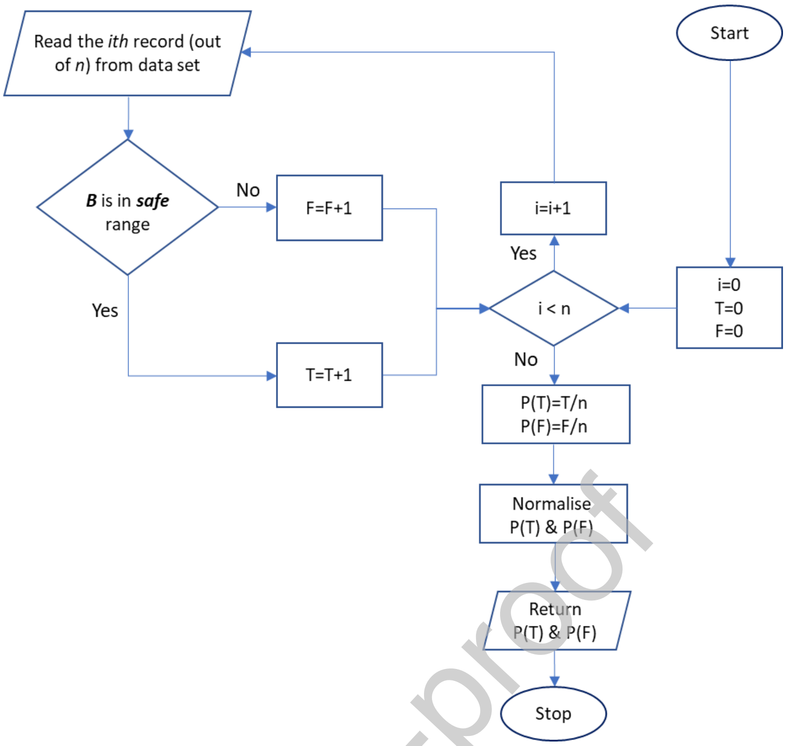



**Figure 5.** The algorithm of extracting initial probabilities for variable $B$ with $n$ records

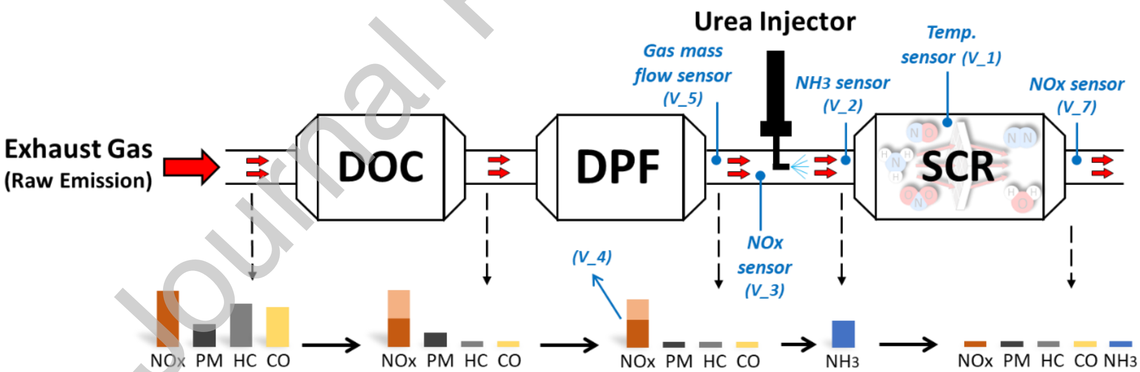



**Figure 6.** A schematic illustration the automotive Aftertreatment system, focused on SCR subsystem analysis

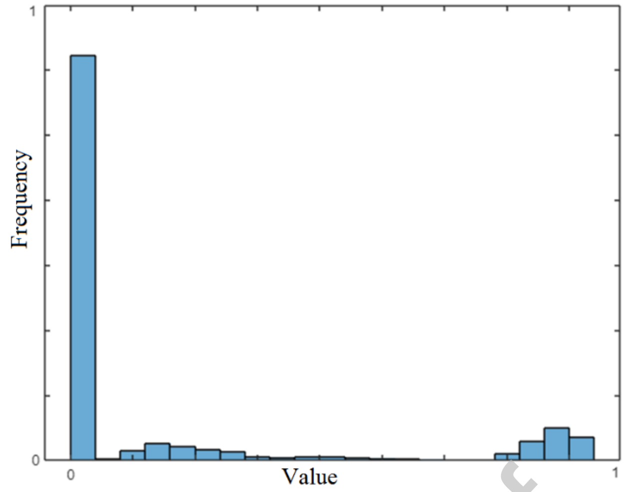



**Figure 7.** Scaled (0 to 1) data for normal operating condition of SCR system

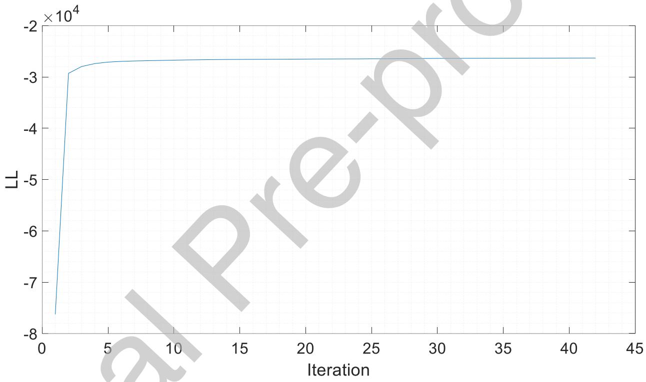



**Figure 8.** HMM training curve

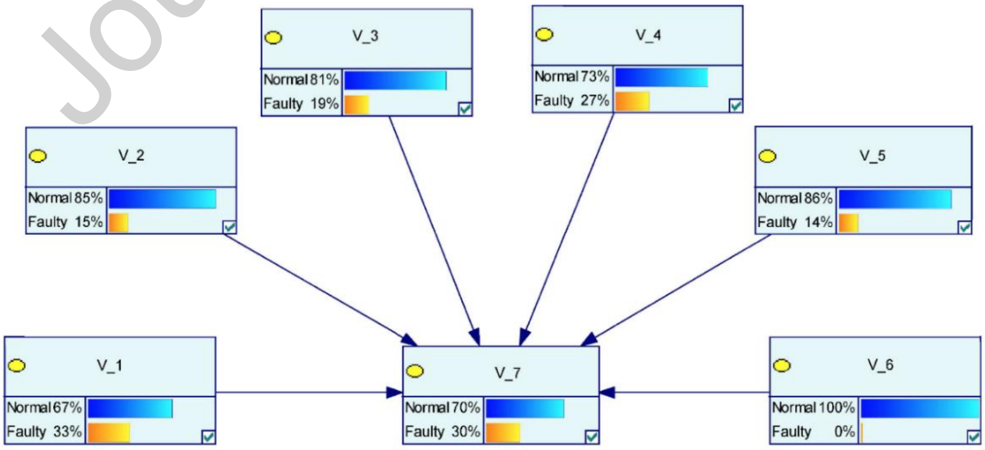


**Figure 9.** BN for SCR system

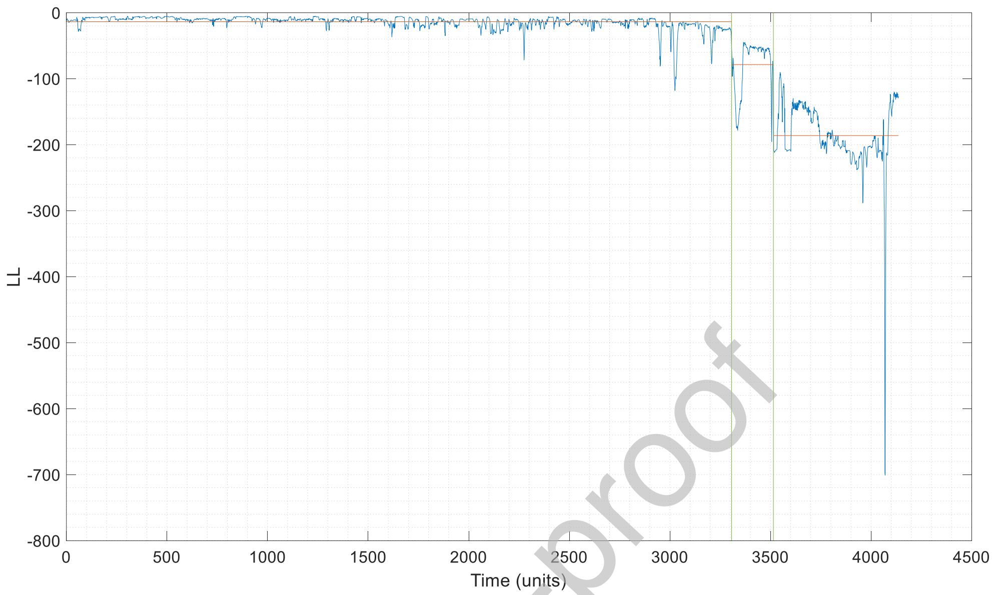



**Figure 10.** Detection of a fault for the new data set

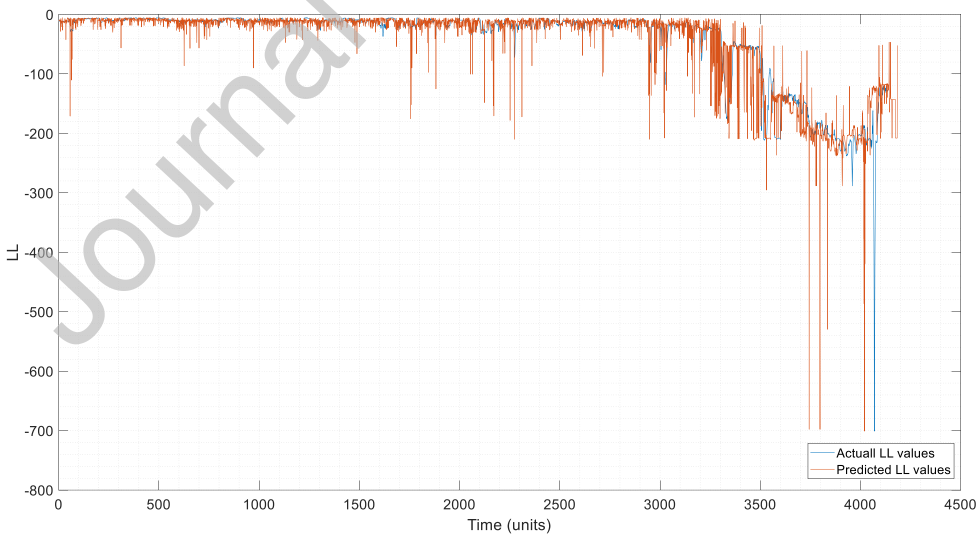



**Figure 11.** Comparison of the actual and predicted LL values for the new data set

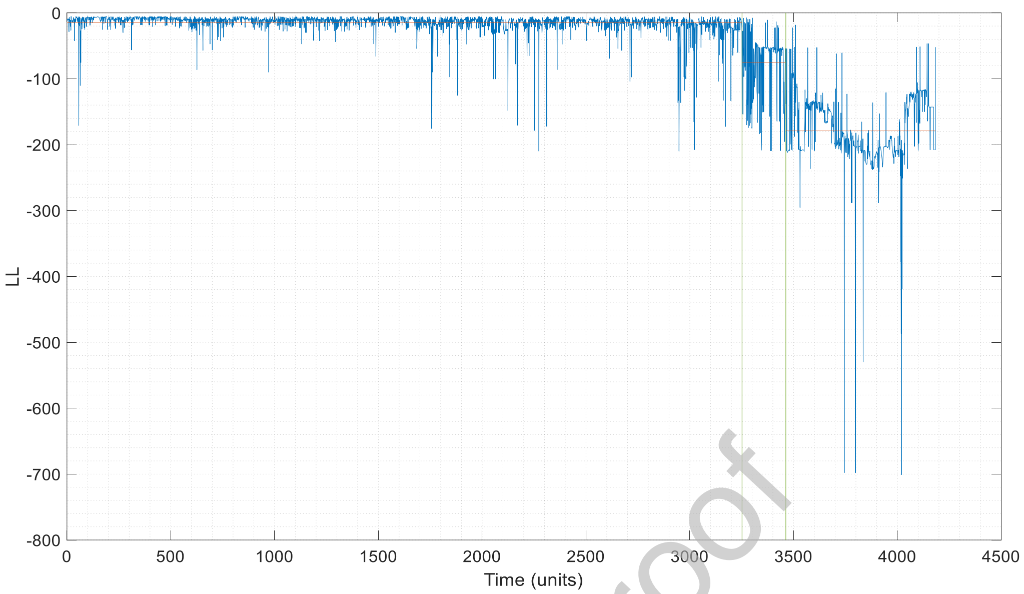



**Figure 12.** Prediction of the fault for the new data set