Open your mind. LUT.
Lappeenranta University of Technology

# Automatic Classification of the Operating Efficiency of Centrifugal Pumps
# Keskipakopumppujen toimintatehokkuuden automaattinen luokittelu

Harri Rahikainen

Bachelor's Thesis
LUT School of Energy Systems
Electrical Engineering

11.11.2017

## ABSTRACT

Lappeenranta University of Technology
LUT School of Energy Systems
Electrical Engineering

Harri Rahikainen
**Automatic Classification of the Operating Efficiency of Centrifugal Pumps**

2017

Bachelor's Thesis.
25 p.

Director & examiner: D.Sc. Tero Ahonen

Pumping systems account for almost a fifth of global motor electricity consumption. With large consumption, system efficiency becomes increasingly important. Efficiency monitoring is typically handled by observing key process variables or machinery and performing separate energy audits. By applying methods of automatically identifying pump efficiency to gathered process data, savings on expensive auditing due to instrumentation costs could be made.

In this Bachelor's Thesis, methods for automatically identifying centrifugal pump operating state are studied. A literature review is performed on the existing research material concerning pump efficiency monitoring. Commercially available solutions are examined. Studied methods and process monitoring are discussed in the case of motor phase current data from industrial pumps. Sample features are extracted from the current data and tested with selected machine learning algorithms. Suggestions for further research and development are proposed.

Commonly used classification methods for pumping applications were reviewed and their basic principles were explained with examples from previous research. Overview of commercial applications found existing products insufficient in terms of automatic classification and more geared towards upgrading traditional fixed-speed systems. Available current data was examined for features and the workflow for doing this was described. Extracted features were tested in MATLAB (MathWorks) environment with six supervised machine learning algorithms, an artificial neural network and a self-organizing map. The reliability of test results is weakened by the lack of sufficient available labelled test data. Systematic collection of more pump operating state data for labelling and further feature engineering were identified as the key focus areas of possible future research.

Keywords: centrifugal pump, energy efficiency, automatic classification, efficiency monitoring

**TIIVISTELMÄ**

Lappeenrannan teknillinen yliopisto
LUT School of Energy Systems
Sähkötekniikka

Harri Rahikainen
**Keskipakopumppujen toimintatehokkuuden automaattinen luokittelu**

2017

Kandidaatintyö
25 s.

Ohjaaja & tarkastaja: TkT Tero Ahonen

Pumppausjärjestelmät käsittävät lähes viidesosan globaalista moottoreiden energiankulutuksesta. Suuren kulutuksen myötä järjestelmien tehokkuuden merkitys kasvaa. Tehokkuuden valvontaa suoritetaan tyypillisesti tarkkailemalla prosessin tärkeimpiä muuttujia tai laitteita ja suorittamalla erillisiä energia-auditointeja. Soveltamalla pumppujen tehokkuuden automaattisen tunnistamisen menetelmiä kerättävään prosessidataan voitaisiin tulevaisuudessa saavuttaa säästöjä mittalaitteiden ja niiden asennusten vuoksi kalliissa auditoinneissa.

Tässä kandidaatintyössä tarkastellaan keskipakopumppujen toimintatilan automaattisen tunnistamisen menetelmiä. Työssä suoritetaan kirjallisuuskatsaus olemassa olevalle tutkimukselle koskien pumppujen toimintatehokkuuden tarkkailemista. Lisäksi tarkastellaan kaupallisesti saatavilla olevia ratkaisuja sekä tutkittuja menetelmiä ja prosessinvalvontaa teollisuuspumppujen moottorivirtadatan tapauksessa. Analyysissä erotetaan virtamittausdatasta esimerkkipiirteitä, joita testataan valituilla koneoppimisalgoritmeilla. Lopuksi annetaan ehdotuksia jatkotutkimusta ja -kehitystä varten.

Tavallisesti pumppausjärjestelmille käytettyjä luokittelumenetelmiä käytiin läpi ja niiden perusperiaatteita selitettiin aiemmasta tutkimuskirjallisuudesta poimittujen esimerkkien kanssa. Kaupallisten sovellusten katsauksen perusteella voidaan todeta nykyisten tuotteiden olevan automaattisen luokittelun osalta riittämättömiä ja suunnattuja pikemminkin perinteisten vakionopeusjärjestelmien päivittämiseen. Käytettyä virtadataa tarkasteltiin esimerkkipiirteiden löytämiseksi ja tämän työnkulku kuvattiin. Löydettyjä piirteitä testattiin MATLAB (MathWorks) ympäristössä kuudella valvotulla koneoppimisalgoritmilla, yhdellä neuroverkolla sekä itseorganisoivalla kartalla. Testitulosten luotettavuutta heikentää käytettävissä ollut riittämätön määrä luokiteltua testidataa. Jatkotutkimuksen pääkohteiksi tunnistettiin piirteiden jatkokehitys ja pumppujen toimintatiladatan systemaattinen kerääminen luokittelua varten.

Avainsanat: keskipakopumppu, energiatehokkuus, automaattinen luokittelu, energiatehokkuuden seuranta

**CONTENTS**

## LIST OF ABBREVIATIONS AND SYMBOLS

| | |
|---|---|
| ANN | Artificial Neural Network |
| ANSI | American National Standard Institute |
| BEP | Best Efficiency Point |
| BMU | Best Matching Unit |
| EFEU | Efficient Energy Use research program |
| HI | Hydraulic Institute |
| KNN | K-Nearest Neighbors |
| MLP | Multi-Layer Perceptron |
| PCA | Principal Component Analysis |
| POR | Preferred Operating Region |
| SVM | Support Vector Machine |
| VSD | Variable Speed Drive |

| | |
|---|---|
| $H$ | head |
| $P$ | power |
| $Q$ | flow rate |

| | |
|---|---|
| $E_s$ | specific energy consumption |
| $g$ | standard gravity value |
| $\eta$ | efficiency |
| $\rho$ | fluid density |

**Subscripts**

| | |
|---|---|
| d | dynamic |
| in | input |
| rel | relative to nominal value |
| st | static |
| tot | total |

# 1. INTRODUCTION

Pumping systems in the industrial sector consume 19 % of global motor electricity demand (IEA 2011). Due to the systems being heavy on energy, it is important that they function efficiently. However, this is often not the case. Pumps are designed for specific process or load conditions, but in reality, their typical operating conditions may vary outside this design point or range. Reasons for these variations may include overestimating the expected load, change in the process variables or mechanical faults and wear. Operation outside the preferred operating range (POR) of the pump can result in a significant drop in the pump's energy efficiency.

Inspection of the pump's condition is traditionally done via scheduled audits or maintenance procedures. Bigger processes can feature on-line monitoring systems for keeping track of the process variables. Audits and monitoring instruments are often expensive and therefore are usually only targeted at the most crucial pumps in the system. Simultaneously, a large number of smaller pumps can run inefficiently. This leads to hidden losses in efficiency and increases the risk of machinery breakdown, when performance degrading faults or wear go unnoticed.

The objective of this thesis is to study methods of automatically identifying the efficiency of centrifugal pumps in their given operating state. For this purpose, several types of automatic decision-making algorithms are studied and evaluated. In the case of energy audits involving a large number of pumps, the costs of instrumentation rise. In order to assess this problem, the identification process should require a minimum number of sensors for data collection. In this thesis, the variables provided to the algorithms are extracted from phase current data collected from the motor driving the pump as normal process measurements. Ideally, an application utilizing these types of algorithms would use the already existing process data and give the user an indication whether pumps are running at good efficiency or not.

This thesis has been conducted under the Efficient Energy Use research program (EFEU), in which the author's university is a partner. The EFEU program, consisting of 11 industrial partners and 5 research organizations, aims to develop energy efficient solutions and services according to its four main objectives. These objectives include the development of measurement, analysis and optimization methods at the system level, applying these methods to applications in fluid handling and regional energy systems and assessing possible business opportunities, collaborations and future developments related to energy efficiency (CLIC Innovation 2016).

The following sections first introduce the studied algorithm types and existing commercial solutions regarding pump efficiency monitoring. In section three, available data are analysed for input variables to the algorithms. Lastly, the algorithms are tested in section four and conclusions are given in section five.

## 2. AVAILABLE METHODS AND SOLUTIONS

The efficiency of pumping systems is affected by the system layout and device characteristics. The system layout includes head losses due to static and dynamic head ($H_{st}$, $H_d$), in other words, overcoming of elevation and friction in the surrounding piping. An example of a system with static head losses is presented in Fig. 2.1 (Ahonen et al. 2015). The pump drive consists of a pump-driving electric motor and the pump itself, each with their unique efficiency values. Factors related to the pump itself include mechanical losses due to, for example, the bearing frame, stuffing box and mechanical seals as well as hydraulic losses occurring through friction and volumetric losses through wear rings (Ferman et al. 2008).

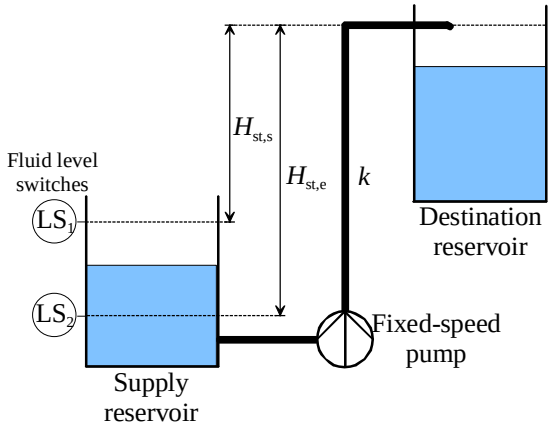


Fig. 2.1. A pumping system consisting of two tanks and a fixed-speed pump. Elevation causes static head loss $H_{st}$ in the process.

Pump characteristics are presented with pump curves provided by the manufacturer. Pump curves indicate the flow rate that a pump can deliver at certain head levels. Curves typically contain total head or power consumption graphed against a range of flows (*QH* and *QP* curves). Curves are usually provided for multiple different pump rotational speeds and impeller diameters. The system requirements are depicted as the total head required to overcome elevational changes and hydraulic friction in the system. The system curve shows the total head needed to move fluid through the system at a given rate of flow. Exemplary pump and system curves are shown in Fig. 2.2 (Grundfos 2017a).

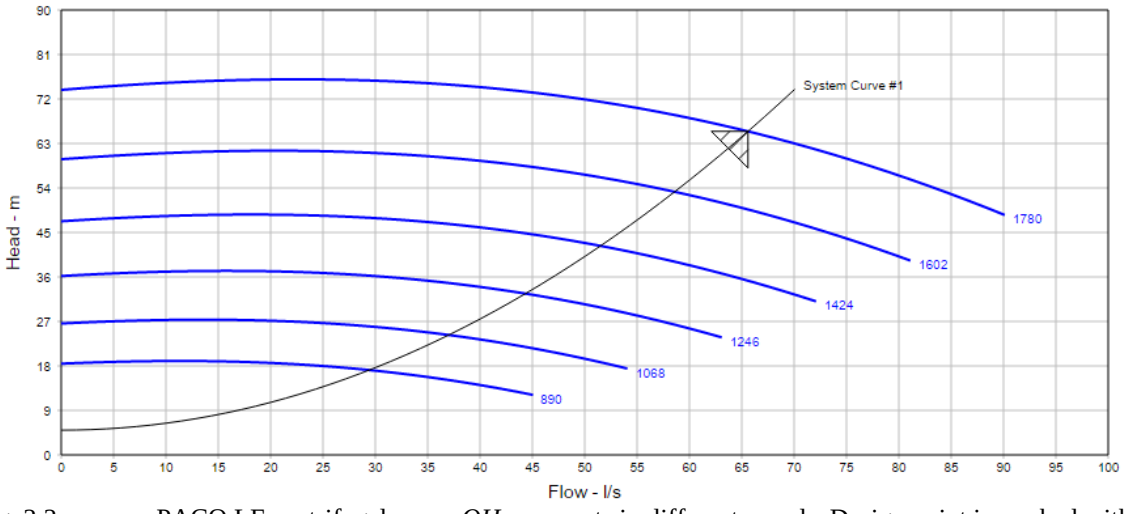


Fig. 2.2. PACO LF centrifugal pump *QH* curves at six different speeds. Design point is marked with an arrow pointing at the intersection of the system and highest speed pump curve.

System inefficiencies are a sum of all the components, therefore, pump inefficiencies are linked to poor system design. Pumps operate in the intersection of the pump and system curves. Ideally, this operating point should be designed and maintained near the best efficiency point (BEP) i.e. the flow rate, which produces the highest efficiency (Ferman et al. 2008).

Operation away from the BEP always reduces the efficiency of the pump. Inefficiencies typically occur for three main reasons. Firstly, system designers tend to leave a safety margin for estimating the system curve incorrectly or for future process changes, which leads to oversizing the motor and the pump. In this scenario, the motor load stays far below optimal, and the drive as a whole runs inefficiently. Another side of misdimensioning is that while running in its BEP, the pump can produce excess flow to the system. This may, in turn, be regulated with throttling, which moves the operating point away from the BEP and decreases the pump's efficiency. Secondly, the system can be poorly operated. This includes throttling, unnecessary continuous operation and fluid recirculation. Finally, most issues come down to the lack of accurate information about the intended and current operating status. In order to make correct decisions in terms of initial design as well as component maintenance, accurate and regular measurements are needed.

While improvements to system efficiency can be made by improving system design to better meet the requirements of the system and applying appropriate system control methods, this thesis focuses on detecting overall pump efficiency through acquiring and analysing measurement data. The effects of assessing individual component inefficiencies are discussed in papers by Maaranen (2010) and Ahonen (2015).

## 2.1 Former research and types of classification

In this section, multiple different types of classification methods are introduced in addition to the studied commercial solutions.

### 2.1.1 Crisp limit rules

The simplest and most straightforward way of performing classification is defining a set of rules, which dictate the appropriate assignment or action for each input. Defining these rules requires detailed knowledge of each process and the variables associated with it. Rules can be presented as crisp IF-THEN statements evaluating each input simply as true or false for each step or class border.

For example, rules can be set for process variables and their boundary conditions. A range-based condition could be set so that operating points outside of the POR of 70-120 % of nominal flow would be classified as inefficient. The limits here were acquired from the POR definition in Hydraulic Institute's standard (ANSI/HI 1997). Rules could also be set for a variety of other variables, but it can be seen that this approach would require a lot of process-specific knowledge. Maintaining a large rule-base would also quickly become cumbersome. Considering these weaknesses, it can be seen how this method of classifying pump performance lacks the generalization ability needed for working with scarcely available data and a large set of pumps.

### 2.1.2 Fuzzy rule sets

Enhancements to a rule-based system can be made by applying fuzzy logic. In contrast to binary logical decisions with Boolean operators, fuzzy systems are suited for describing the gradual nature of real-life processes. Instead of crisp value limits, fuzzy logic introduces a degree of belonging to a class. In a fuzzy system, inputs and outputs are given linguistic terms to describe their numerical value. In the case of a pumping system, relative flow rate and specific energy consumption could be assigned terms low, medium and high and the pump condition would similarly be either highly inefficient, slightly inefficient or efficient. Changes from one term to another are defined by the so-called membership functions. Membership functions indicate the degree of the variable being described with a specific linguistic term with values between [0, 1]. The shape of a membership function has to be considered based on the type of data it represents. A simple triangular shape, as shown in Fig. 2.3, can be appropriate for variables that have well-defined boundaries between terms. More complicated data might use a Gaussian or a highly custom function shape.

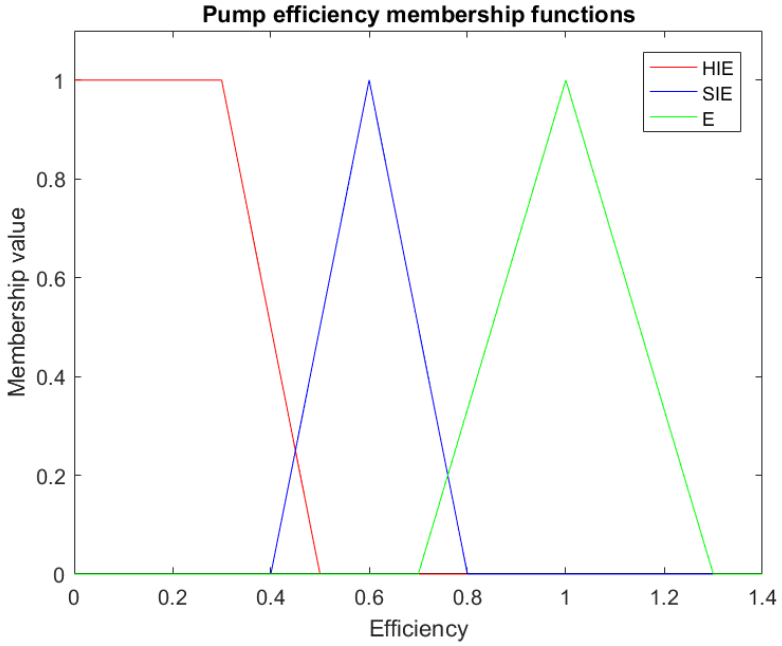


Fig 2.3. Triangular membership functions for pump efficiency. Exemplary conditions interpreted as highly inefficient “HIE”, slightly inefficient “SIE” and efficient “E”.

The output of a fuzzy system is determined based on IF-THEN rules formed via expert knowledge of the system. In contrast to rough rule-based systems, fuzzy systems use the linguistic terms dictated by the membership functions instead of raw numerical values. Rules follow the format “If variable A is of linguistic condition B, then output is of condition C.”. After evaluating the rules that were fired, the overall output is determined from the combination of affected output memberships and the output condition is defuzzified back to a numerical value. A fault detection system using fuzzy logic is presented by Rodriguez and Arkkio (2008).

Fuzzy rules are easy for humans to understand, due to their linguistic presentation. This makes them viable for use in control systems, where supervisor’s knowledge can be translated to control rules in a natural way. A downside to the fuzzy approach is that knowledge and deep analysis of the available data is required to form fitting membership

functions and rules. Unforeseen patterns in the data may stay unnoticed if they aren't governed by rules explicitly. A combination of rough sets and fuzzy logic for fault diagnosis is introduced in Muralidharan and Sugumaran (2013).

### 2.1.3 Machine learning algorithms

In situations, where the exact causalities between process variables are unknown, it can be useful to utilize machine learning. Machine learning algorithms can help uncover the underlying functions and nonlinear relationships in the data. Machine learning is a broad term, but for the purposes of this thesis, it is used to refer to both the field itself and its sub-field data mining.

For determining pump efficiency, two types of machine learning were studied in particular: supervised and unsupervised learning. In supervised learning, the computer is given input data and the corresponding target outputs and is trained this way to form a rule to map these together. With unsupervised learning, no target values are given, and the computer has to find structures in the data by itself. Data mining focuses more on revealing unknown properties from the data and can be considered as a form of unsupervised learning. Semi-supervised learning can be effective on large datasets where only a small number of target values/labels are known (Alpaydin 2014).

In terms of desired output from the learning process, two tasks were considered: classification and clustering. In classification, the model is trained with inputs labelled to different classes for the model to correctly assign completely unseen inputs to their corresponding class. This is usually done via supervised learning methods by assigning to each training data input their desired output. A feature refers to an attribute identified from the input variables. As a result of training, the dataset is organized based on similarity in some chosen metric between the input feature values. Using supervised methods, classification is viable, when desired groupings are known beforehand. Classification is essentially an optimization problem, where the difference between desired output and given input is calculated by a loss or error function, which the algorithm then tries to minimize. Clustering is structuring input data into groups based on similarity. However, this is usually done in an unsupervised manner and the number and qualities of groupings are not initially known. When presented with unfamiliar data, it can be useful to first run an unsupervised clustering algorithm and use a supervised algorithm, once classes can be identified from clusters (Alpaydin 2014).

During the background research for this thesis, some widely used algorithms for classification problems were studied. Descriptions of each algorithm's main principles are presented here. Algorithms best applicable for pump classification are discussed in more detail in section three.

**Artificial Neural Networks (ANN)**, inspired by the functionality of the human brain, are widely used algorithms in classification and fault detection problems. Neural networks consist of layers of perceptrons and weights connecting them. Networks with one or more additional layers between the input and output layer are called Multi-Layer Perceptrons (MLP). MLPs can be used to perform nonlinear discrimination on the data (Alpaydin 2014).

The input layer consists of input perceptrons (dimensions) and a bias perceptron with a value of one. Each layer after the input layer takes the preceding layer values as input and passes them though a nonlinear or radial basis function such as the sigmoid or hyperbolic tangent

(Gaussian). Finally, the outputs are calculated as linear combinations of the nonlinear basis function values. The weights are then updated based on the difference to the desired output. This is often done via backpropagation algorithm proposed by Rumelhart et al. (1986). The algorithm updates the weights starting from the last hidden layer by minimizing an error function. The method is called Gradient Descent and attempts to optimize weights so a minimum is reached. Backpropagation can get stuck on local minima, which slows convergence. Another parameter to carefully consider is the learning rate, which essentially dictates how large the weight updates will be each time. A comparison of logistic regression, support vector machines and multi-layer ANNs found MLPs based on principal component analysis (PCA) more effective in monitoring oil pump suction strainer health. Although computationally more expensive, ANNs proved superior in generalization ability (Raza et al. 2010).

**Self-organizing map (SOM)** is an unsupervised method of learning proposed by Kohonen (1990). Self-organizing maps are especially usable in problems where it is necessary to process high-dimensional data. They perform a dimensional reduction on the data, usually to 2D/3D, making visualization of the data much easier.

The maps are variations of ANNs, which consist of cells, neurons. The variables or attributes of a given problem are presented as weight vectors assigned to each neuron at the beginning of the algorithm. Weight vectors can be initialized by random or through methods like PCA. The map is then trained by inputting data of the system. The learning is achieved by determining the neuron that best represents the input data. This neuron is referred to as the best matching unit (BMU). An often-used metric is the Euclidean distance. After a BMU has been selected, it is then moved closer to the input vector, making it even more similar in terms of its attributes. The neurons close to the BMU, in its so-called neighbourhood, are also shifted towards the BMU, but to a lesser extent. The functions related to this process are the neighbourhood function and the learning rate, which dictates the distance shifted. Both of these start larger and decline over time so that through longer series of iterations, eventually, only the BMU is affected. Continuing this process maps the input vectors to their corresponding neurons on the map while preserving the topology (Kohonen 1990).

Through training the neurons become sensitive to certain types of attributes and start to form denser areas of similar neurons, clusters. Now additional input data, test points, can be used to identify clusters containing certain attribute combinations i.e. classes. By assigning the test vectors appropriate values based on prior knowledge it is possible to approximate the class borders. An example of SOMs in industrial process monitoring is presented in Dominguez et al. (2007). The application is based on tracking the BMU and detecting deviations from normal operation trajectory. A labeling method by Rauber (1999) also seems promising. This method called LabelSOM attaches labels to each neuron on the map according to the features that contribute most to an input vector landing on that neuron.

**Decision trees** split the input space into two or more branches i.e. subsets at each decision node until all inputs following each branch belong to the same class. The probability of ending up on a specific leaf node, in other words class, is then the ratio of the instances on the leaf to the number of instances reaching the preceding decision node. Split decisions are made based on an impurity measure. A split is pure if for all the new branches, all the inputs choosing the same branch are of the same class (Alpaydin 2014). Splits that decrease impurity lead to smaller trees needed to differentiate classes. Univariate trees use only one input dimension to make their decisions, whereas multivariate trees can utilize multiple

dimensions at each decision node. To avoid noise causing the classifier to overfit, pruning can be done to improve the generalization ability of the tree. This can be achieved by stopping the tree-growing if too few instances support the decision or by removing redundant rules after growing the tree. The advantages of decision trees are interpretability, simultaneous feature extraction and simplicity. After the tree has been constructed, it can be easily read as rules. Furthermore, by examining the decision nodes one can distinguish features in the data that affect classification the most. An example of using a decision tree for pump fault detection is presented in Sakthivel et al. (2010).

**Support Vector Machines (SVM)**, proposed by Vapnik (1995), work by fitting a hyperplane into the input space in a manner that minimizes the number of incorrect classifications. The hyperplane is positioned so that it is a maximum distance away from the members of each class at any given point. The hyperplane serves as a decision boundary between classes. SVM hyperparameters and kernel functions need to be correctly chosen for good results. Vibration and wavelet based fault diagnosis of centrifugal pumps using SVM is discussed in Muralidharan et al. (2014). Other discriminant-based methods include for example linear, quadratic and cubic discriminants, which have decision boundaries of different shapes (Alpaydin 2014).

**K-Nearest Neighbors (KNN)** algorithm is a simple method for classification. The method is based on determining the class of an input vector from known classes of its *k* nearest vectors in the training set. A special case of this is k = 1, where input are assigned to the single nearest neighbor's class. Variations of the algorithm may use weights based on the distance of the neighbors, for example, a weight equal to the inverse of the distance to the input. The distance metric used needs to be considered depending on the type of data (Alpaydin 2014).

## 2.2 Commercial solutions

For comparison with the scientifically researched classification and decision-making methods, available commercial market applications were studied. The amount of supervision needed and the ability to apply the solution to a large sample of pumps were the main points of investigation. Most relevant applications are listed in Table 2.1.

Table 2.1. Researched commercial applications for pump monitoring (ABB 2010, ABB 2013, ITT Water and Wastewater 2017, Siemens AG 2014, Grundfos 2017b).

| Application name | Service provider | Features and primary use |
|---|---|---|
| PEMS-Pump Efficiency Metering System | ABB | Thermodynamic measuring method: 3 measurements for efficiency monitoring, predictive maintenance recommendation, integration with controller handles sets of pumps |
| PumpFit | ABB | Start/Stop control, pump selection for lowest overall consumption, limit and operating hours monitoring |
| Flygt Smart Compact: Control Panel | Flygt | Integrated VSD, using PumpSmart control, sensorless flow monitoring based on characteristic curve information and power and torque measurements, multi-pump load balancing |
| PumpMon | Siemens | A part of the SIMATIC PCS 7 Condition monitoring library, 5 measurements, limit and tolerance monitoring of characteristics |
| CR Monitor | Grundfos | Control panel for CR pumps, self-configuring with 6 measurements, comparison to measured or set reference levels |

The investigated products focus largely on improving system efficiency via means such as variable speed drives (VSDs) and incorporating system level changes. There exists a plethora of applications such as the Flygt “Smart Compact Control Panel” that offer extensive measurement capability and monitoring system utilities geared toward new installations or system upgrades. However, these systems are rarely independent in decision-making and need a supervisor to make the conclusions out of the presented data. Another large market share is taken by variable speed drive and frequency converter solutions to upgrade traditional fixed speed drives. It should be noted that manufacturers provide little information about the actual science behind their control methods and most encountered material is marketing material. Also, most of the studied “intelligent” monitoring systems required several sensors for measuring the operating status, so there seems to be a demand for a non-intrusive monitoring method such as the one discussed in this thesis.

# 3. DATA AND EXTRACTED CHARACTERISTICS

The focus of this thesis was to study methods for determining whether centrifugal pumps are functioning efficiently based on process measurements. This thesis uses data from a prior master's thesis by Maaranen (2010). In his thesis, Maaranen discusses the energy efficiency of pump drives and how it could be improved using case examples. These examples are based on analysing pumps in a cardboard machine system in Kaukopää mill, Imatra, Finland. Maaranen utilizes phase current measurements from the pump-driving motors and the characteristic curves of the corresponding pumps to provide estimations for motor shaft power, pump operating point and efficiency. He then proceeds to evaluate the efficiency of the pump drive based on a set of criteria. In total, 37 drives were studied, 18 of which were VSDs.

The measurement data used were collected during a three-month period between 11/4/2009 and 4/2/2010. Phase current data are averaged with a time interval of 10 minutes. An example of current data is presented in Fig. 3.1. Motor shaft power is approximated with a second order polynomial fitted into current and power values given at partial loads by the manufacturer. Points from the characteristic curves were digitized and shaft power estimates were used to solve flow rate values from the *QP* curve. Flow rates were in turn used for solving head values from the *QH* curve. Estimated values were used to calculate pump efficiency using cold water as the sample fluid. Linear interpolation was used to solve points on the characteristic curves. Estimating values based on the characteristic curves introduces some uncertainty, considering that the real operation may differ from the published curve, as discussed in Ahonen (2011). Furthermore, the estimation algorithm doesn't correctly handle certain cases, where the characteristic curve doesn't increase linearly, which leads to the algorithm linearizing between subsequent points and missing the real curve trace. An example of estimation failure is shown in Fig. 3.2. On flow rates relatively close to the BEP, however, the estimation works sufficiently.

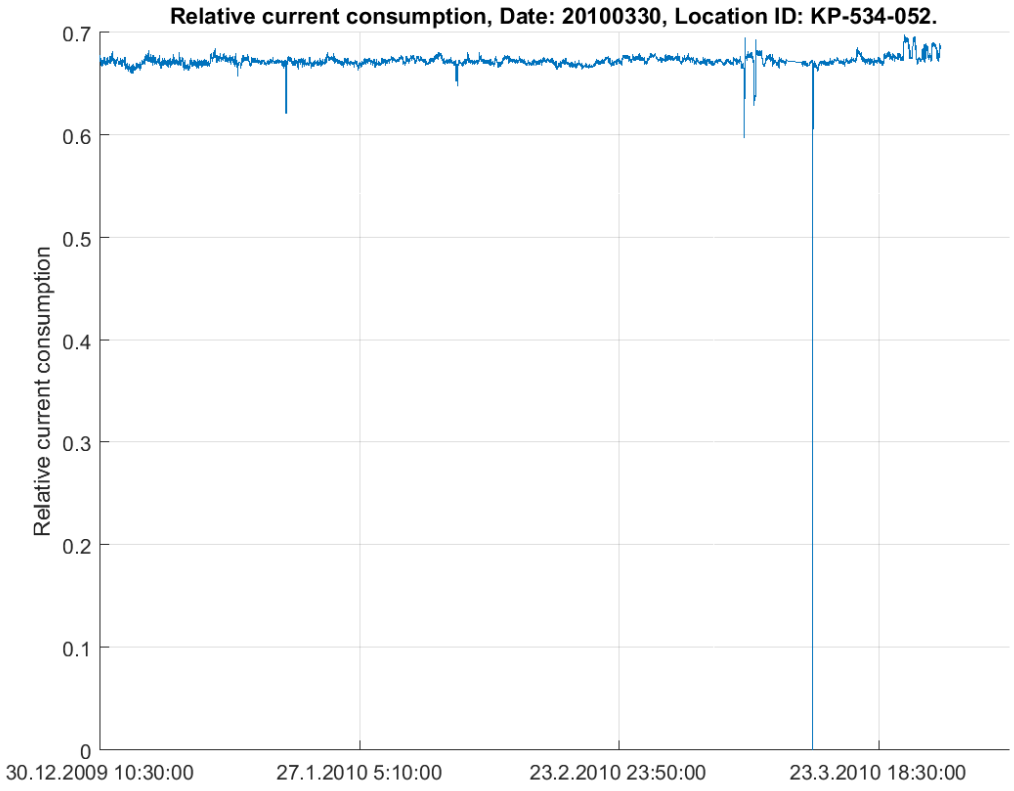


Fig. 3.1. Relative current data measurements from pump KP-534-052. Zero value spikes indicate maintenance procedures.

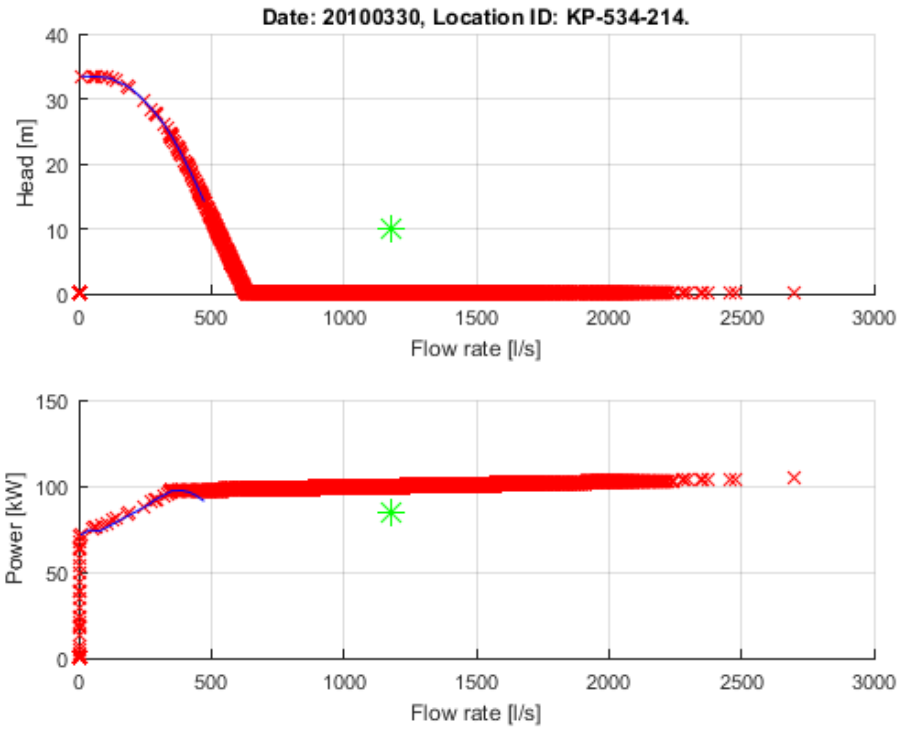


Fig. 3.2. Failed estimation results. Manufacturer's curve is presented in blue, estimated operating points in red. Green asterisk represents the mean of operating points. Estimates that produced negative flow rates are positioned at zero flow for simplicity.

The criteria for pumps set by Maaranen divides studied pumps into three categories or classes. The first class consists of pumps functioning near their nominal operating point. The POR described in the thesis for relative flow rate is 80-120% of nominal flow rate. The Hydraulic Institute (HI) standard is 70-120 % (ANSI/HI 1997). Pumps of the first class are considered efficient. The second class includes pumps that operate at relative flow rates of under 80% or over 120% of nominal flow. These pumps are operating outside of their POR, at partial flows. The third class holds pumps, the utilization rate of which is small or for which the estimation method produced erroneous results. The inspection of drive efficiencies also considered motor relative load when determining improvement options. Although crisp, the criteria set by Maaranen can serve as a reference for evaluating more elaborate classification methods.

### 3.1 Pre-processing and features

Classification and clustering are based on finding similarities in attributes of the data. Attributes, commonly called features, are characteristics, which contribute to the output and are calculated or deduced from the data. Features need to be descriptive of the target class. As such, features should have either high relevance in respect to the target or both high relevance and minimum redundancy i.e. correlation between features. This ensures that each computed feature gives new information. In the case of efficiency and/or fault detection, causalities between a feature value and the output of the classification need to be determined.

Dimensionality reduction methods are employed to reduce the number of features needed to compute and input to the monitoring system. Two main methods of doing this are feature selection and feature extraction. In the former, a subset of the available features is selected so that it contains the most information. This can be accomplished by adding or removing features and examining their effect on the classification error. When using the latter method, the aim is to derive new variables from the original set. In principal component analysis, the data points are projected onto principal directions i.e. the directions in which they have the most variance. The number of components that retains the most variance desirable is chosen for use in classification.

Used dataset needs to be split into three parts: the training data set, validation set and the test set, which is introduced to the model as unseen data to classify. In Selak et al. (2014), measured data from a hydropower plant was divided by operating regime to four groups: all data, all but system stops and water levels less than a limit, all former and high water flow removed and finally transient phases removed. The more specific groups allow for earlier detection of abnormalities. It should be noted that the data acquired from the cardboard machine system are averaged current data with a time interval of 10 min. Therefore, the data in question can't be used to detect fast fluctuations. For determining short-term faults, the sampling rate should be adjusted accordingly. Also, the transient phase data need to be filtered in this case so that the 10-min average can be used to recognize normal steady state operation abnormalities.

**Features** to consider in detecting pump efficiency could include, for example, the average duty point of the pump (relative flow value $Q_{rel}$), specific energy consumption $E_s$ (kWh/m$^3$), average efficiency, average current load and utilization rate.

A pump can be run within its POR inefficiently with partial loads. Specific energy consumption, decreasing with improving efficiency, is defined by equation

$$E_s = \frac{P_{in}}{Q} = \frac{\rho g H_{tot}}{\eta}, \tag{3.1}$$

where $\rho$ represents fluid density, $g$ the standard gravity value, $H$ the total head consisting of the sum of static and dynamic head values $H_{st}$ and $H_d$ and $\eta$ the combined efficiency of the pump device and drive train (Ahonen 2015).

Pump efficiency at a given operating point should be assessed relative to the maximum attainable efficiency for the pump. A pump, which is run with high efficiency but is rarely in operation due to its task, is still considered efficient in this thesis. Utilization rate is approximated with the ratio of operating points with non-zero current values to the number of measurements taken. Average motor current is calculated as the mean of non-zero points. Relative values are calculated from flow, head and power estimates by dividing them by nominal values. The set of pumps chosen for further study is presented in Table 3.1.

Table 3.1. Operating characteristics of the chosen pump set.

| Pump ID | Average relative flow rate (%) | Average relative efficiency (%) | Average relative motor current (%) | Utilization rate (%) | Nominal point flow rate (l/s) | Nominal point head (m) | Nominal point motor shaft power (kW) | Nominal point pump efficiency (%) |
|---|---|---|---|---|---|---|---|---|
| KP-534-052 | 99 | 99 | 67 | 99 | 103 | 19 | 23 | 84 |
| KP-534-112 | 81 | 96 | 78 | 75 | 250 | 48 | 137 | 87 |
| KP-534-213 | 103 | 91 | 84 | 85 | 1150 | 24 | 319 | 86 |
| KP-534-253 | 29 | 46 | 84 | 88 | 458 | 35 | 188 | 83 |
| KP-534-271 | 135 | 91 | 94 | 90 | 110 | 53 | 70 | 82 |
| KP-534-618 | 63 | 84 | 42 | 11 | 834 | 49 | 527 | 77 |
| KP-534-675 | 44 | 66 | 50 | 86 | 194 | 33 | 77 | 81 |

For online implementation, a constant evaluation of sample duty point data is needed. Out of the pump data collected by Maaranen, only measurements on which the flow and shaft power estimation worked could be studied. A sample number of under 20 pumps isn't

enough for learning models, therefore additional data were needed. Although models may work correctly on the training and validation sets, testing set results tend to weigh too much toward single feature values (namely $Q_{rel}$). Alternatively, samples from different pumps could be merged, in combination with universally applicable features, to acquire more samples. Results obtained with few samples couldn't make a distinction between pumps operating at partial loads.

## 3.2 Efficiency monitoring workflow

Applying an efficiency monitoring system requires certain steps universal to any used classifying or prediction method. A 4-step exemplary workflow is illustrated in Fig. 3.3.

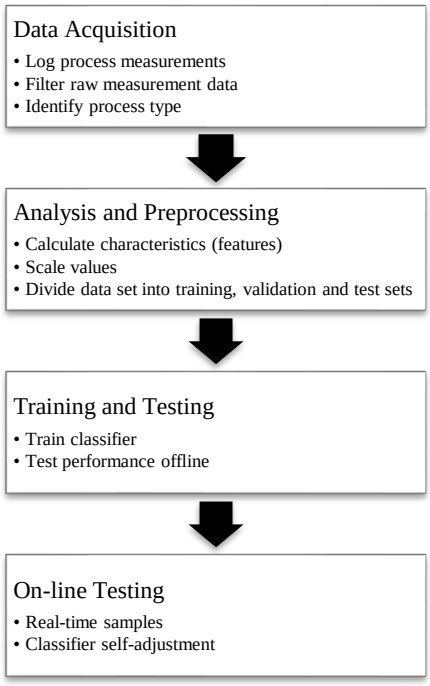


Fig. 3.3. Process efficiency monitoring exemplary workflow

Firstly, data are gathered via process measurements. As discussed in section 1, this should be achieved with the fewest possible sensors and instruments, in order to reduce implementation costs. Frequency converters are capable of providing voltage, current, torque and rotational speed estimates for further processing. After logging the data, an inspection into the form of the data point distribution is necessary to remove outliers and erroneous measurements. If the process type is known, as in the case of tank filling shown in Fig. 2.1, knowledge of its distinctive features could be used to, for example, form fuzzy rules as mentioned in 2.1.2.

Next, filtered data need to be analysed and pre-processed. The calculations may be performed via cloud computing or a system mainframe computer. Once the features are chosen, as discussed in 3.1, common practice is to scale, i.e. normalize or standardize their values for fair comparison. Normalization rescales numeric values in the range [0, 1], whereas standardization gives the dataset zero mean and unit variance. The dataset is then divided into three parts for validation and testing purposes.

The third step is training and testing the classifier. In this phase, the training dataset and validation sets are used to perform the learning and confirm the classification accuracy on the trained model. Testing set performance indicates the model's ability to generalize from unseen data, which is crucial for an on-line implementation that requires constant assessment of new samples.

Lastly, classifier performance needs to be tested in a real-time environment. Learning models need to be configured so that they are less sensitive to outliers and noisy data. An example of this is using an adaptive resonance theory (ART) based network (Srinivasan&Batur 1994). An ART network won't outlearn previously learned patterns, which is usually the problem with on-line implementations.

## 4. TEST RESULTS

Testing in this chapter follows a workflow similar to chapter three. For testing purposes, a combination of 168 samples representing different operating states was chosen from the available measurement data. Samples were chosen so that both efficient and inefficient states were adequately present. Characteristics that were used include relative flow rate, specific energy consumption (both relative and absolute), efficiency and relative motor current, as was discussed in 3.1. Several supervised methods were tested in addition to clustering via a self-organizing map. Testing was conducted using the MATLAB Statistics and Machine Learning Toolbox.

Six supervised methods were chosen for further inspection. These include a decision tree, two K-Nearest Neighbor classifiers, linear and cubic kernel SVMs and a quadratic discriminant classifier. A subset of the 168 samples was put aside for testing. All classifiers managed a 100 percent score on the testing set, which was to be expected since both the training and testing set were from the same hand-picked distribution by the author. In addition, a two-layer ANN with a 10-neuron hidden layer and a softmax layer was tested. The ANN also scored a 100 on the testing set. The confusion matrices of the ANN exhibiting the classification accuracy are presented in Fig. 4.1.


Fig. 4.1. Two-layer ANN confusion matrices for training, validation and test sets. Target classes (labels) are marked 1 and 0 for good and inefficient pumps respectively. Output classes are predictions made by the network. Overall accuracy at each step is presented in the blue square.

The sample set was then constructed into a 10x10 neuron SOM. After 200 iterations of the algorithm, some clusters could be identified. A coloured version of the map is shown in Fig. 4.2.

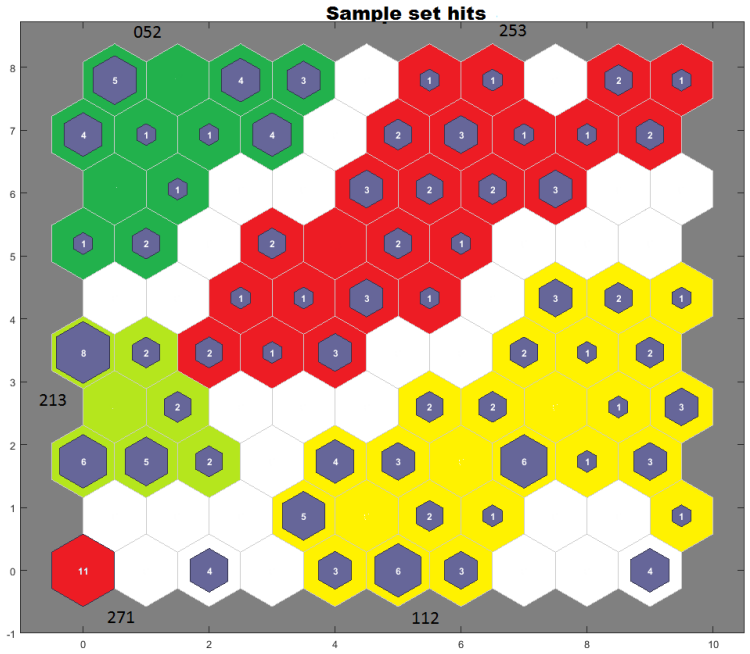


Fig. 4.2. SOM sample set hits. Predicted clusters are coloured in terms of pump operating state, green representing well-functioning and red very inefficient pumps. Numbers next to clusters indicate location IDs of the pumps that contributed most to them.

From the figure, it can be seen that approximately five different clusterings were formed. Known operating points from the sample set were used to identify which neurons on the map reacted to them. Clusters were then manually coloured based on the degree of efficiency of their subset of pumps. Means of the feature values representing each cluster are shown in Table 4.1. On the left side of the map, well-functioning pumps are presented in green. Light green represents pumps with relative values slightly over 1, whereas the darker shaded cluster has values near ideal with little variation. The yellow cluster represents pumps that operate out of their POR, with flows ranging from 60 to 80 percent. The bigger red cluster was formed by operating points where flow rates were low (20-50 %) and, consequently, specific energy consumption very high. The single red neuron represents flow values of over 130 %, with relative specific energy consumption dropping under 1.0. The remaining uncoloured neurons consisted of value combinations not quite belonging to any of the groups, but the likelihood of them belonging to the red and yellow clusters is high.

Table 4.1. Averages of feature values amidst the training vectors belonging to each cluster. Rows represent their respective pumps.

| Pump ID | Relative flow rate | Specific energy consumption | Relative specific energy consumption | Relative pump efficiency | Relative motor current |
|---|---|---|---|---|---|
| KP-534-052 | 1.11 | 0.75 | 0.93 | 0.97 | 0.68 |
| KP-534-112 | 0.73 | 2.37 | 1.20 | 0.92 | 0.76 |
| KP-534-213 | 1.05 | 1.04 | 1.04 | 0.90 | 0.85 |
| KP-534-253 | 0.37 | 3.81 | 2.58 | 0.56 | 0.87 |
| KP-534-271 | 1.41 | 2.02 | 0.88 | 0.87 | 0.96 |
| KP-534-675 | 0.92 | 1.53 | 1.07 | 0.99 | 0.72 |

The same procedure was implemented with the whole base set of some 98,000 samples collected and filtered for estimation errors. Testing with the labelled samples yielded similar results, but with more unexplained variance. The map of the base set is shown in Fig. 4.3. The lone red cluster again represents excessive flow rates. Most of the variation is likely explained by flow rate and specific energy consumption, as in the lower part of the map flow rate decreases from the bottom up. The red cluster is scattered most likely due to high variation in specific energy consumption within the group, relative values ranging from 2.5 to 9. These hypotheses are based on first testing with the known pump operating points and then modifying their input parameters to study the activation of neurons along different dimensions.

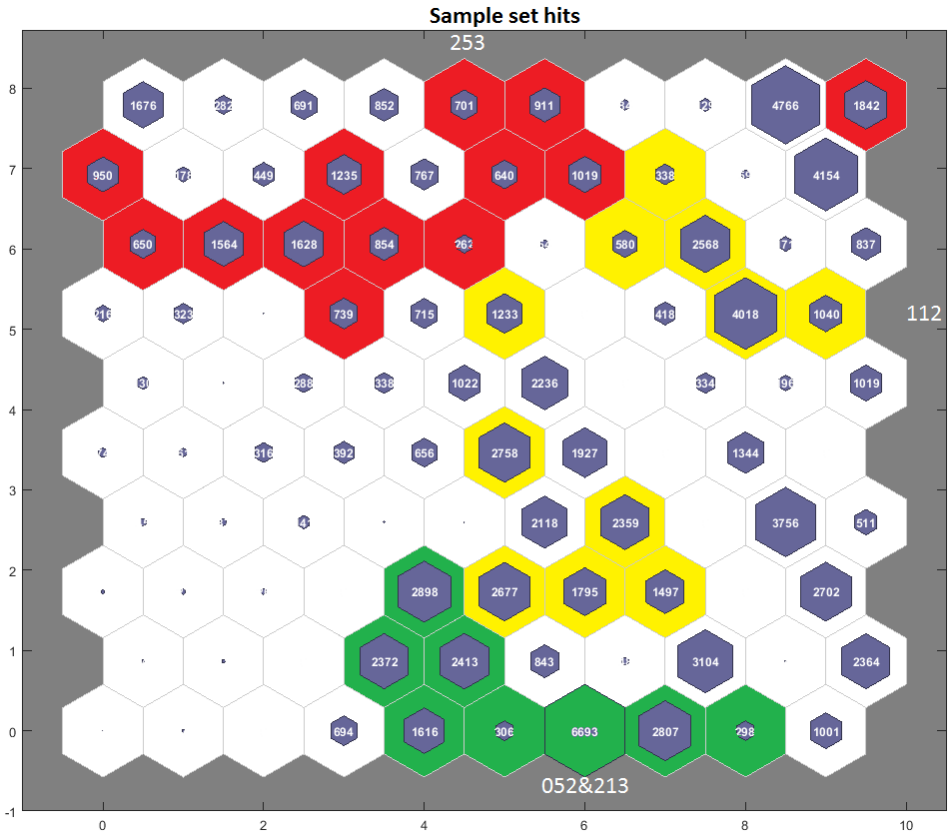


Fig. 4.3. SOM constructed with the base set. Colour coding principles are the same as in Fig. 4.2. Uncertain neurons are left white.

# 5. CONCLUSIONS

The objective of this thesis was to study methods of automatically identifying and classifying the efficiency of centrifugal pumps. A literature review was performed on the existing research material concerning pump efficiency monitoring. Methods applied in the viewed papers were taken under inspection and their basic principles were explained. Commercially available solutions from well-known companies were also examined.

In chapter 3, available data and its collection and properties were studied. The rough criteria for pump efficiency presented by the prior thesis worker were also introduced as a basis for further development. Data preparation and pre-processing were discussed in general and with examples applied to the examined pumps. The relative values of flow rate, specific energy consumption, efficiency and motor current were chosen as features for testing. An exemplary efficiency monitoring workflow was presented to summarize the key points.

Testing on the data was performed by taking a sample of operating points from the pumps in Table 4.1. This sample consisted of operating points where the efficiency could be manually determined based on the pump curves and familiarity with the data. The operating points were labelled as well-functioning, slightly inefficient or very inefficient. Six supervised algorithms along with an ANN network were tested with this sample. All seven of these expectedly classified correctly on the test set, which was the same as the training set due to limitations in estimation reliability and only the small sample being labelled data. The choices for parameter values require further tuning and expertise in the field. The selection of characteristics, as discussed in previous chapters, should be re-evaluated, in addition to gathering fresh data systematically for easier labeling and diversity in operating states. Picking samples by hand inevitably introduces some error that is dependent on the author. By deliberately driving test pumps in different efficiency areas, it could be possible to verify the labeling of data and ensure sufficient coverage of the training samples over possible states during real operation.

Unsupervised methods were tested by constructing a 10x10 Kohonen SOM. The sample operating points were then used to test which neurons on the map reacted to their type of operating state. The map was first trained on the sample set of 168 points and in the second test on the whole base set of some 98,000 samples. In both cases, clusters of operating points with similar efficiencies could be roughly identified. However, the black box nature of the SOM makes it difficult to determine which combinations of features result in a particular neuron reacting to the input.

In conclusion, it can be stated that there exists a need for an automatic classification method or application addressing the problem of sensorless and continuous efficiency monitoring. The subject requires more research as studied commercial products do not satisfy these criteria at the moment. Out of the studied methods the SOM and its variants like the one proposed by Dominguez et al. (2007) seem the most appealing for their visual representation of the operating point trajectory. This kind of representation could be highly informative online. Furthermore, fresh labelled data and additional feature engineering are needed to further develop models for pump efficiency classification.

## REFERENCES


ABB, 2010. *Advanced Pump Efficiency Metering System PEMS*. [Online] Available at: https://library.e.abb.com/public/221678c2a704c76cc1257cd10042c825/DEABB%201547%2010%20en%20_%20Pump%20efficiency%20metering%20system_160310.pdf [Accessed 22 October 2017].

ABB, 2013. *PumpFit Finds the number of running pumps with lowest total electricity consumption*. [Online] Available at: https://search-ext.abb.com/library/Download.aspx?DocumentID=DEABB%201916%2013%20en&LanguageCode=en&DocumentPartId=&Action=Launch [Accessed 22 October 2017].

Ahonen, T., 2011. *Monitoring of centrifugal pump operation by a frequency converter*. Lappeenranta: Lappeenranta University of Technology.

Ahonen, T., 2015. *Energy Efficiency Lecture 7: Energy efficiency in electric-motor-driven systems*, Lappeenranta: Lappeenranta University of Technology.

Ahonen, T., Tamminen, J., Viholainen, J. et al., 2015. Energy efficiency optimizing speed control method for reservoir pumping applications.
*Energy Efficiency*. DOI: 10.1007/s12053-014-9282-6.

Alpaydin, E., 2014. *Introduction to Machine Learning.* 3rd ed. Cambridge, Massachusetts: The MIT Press.

ANSI/HI 9.6.3, *Centrifugal and Vertical Pumps for Allowable Operating Region*, 1997.

CLIC Innovation, 2016. *EFEU CLIC Innovation*. [Online] Available at: http://clicinnovation.fi/activity/efeu [Accessed 14 July 2016].

Dominguez, M. et al., 2007. Internet-based remote supervision of industrial processes using self-organizing maps. *Engineering Applications of Artificial Intelligence*, 20(6), pp. 757-765.

Ferman, R. et al., 2008. *Optimizing Pumping Systems: A Guide to Improved Energy Efficiency, Reliability, and Profitability.* 1st ed. Parsippany, New Jersey: Hydraulic Institute; Pump Systems Matter.

Grundfos, 2017a. *Grundfos Express.* [Online] Available at: http://www.grundfosexpresssuite.com/ [Accessed 26 January 2017].

Grundfos, 2017b. *Grundfos CR Monitor Intelligent Warning.* [Online] Available at: http://www.grundfos.com/products/find-product/cr-cre-monitor.html/Grundfosliterature-1563858.pdf [Accessed 22 October 2017].

International Energy Agency, 2011. *Energy-Efficiency Policy Opportunities for Electric Motor-Driven Systems*. pp. 39-40.

ITT Water and Wastewater, 2017. *Flygt Control Panels – An Introduction to the range.* [Online] Available at: http://www.xylemwatersolutions.com/scs/belgium/en-gb/brands/flygt/Monitoring%20and%20control/Documents/3004426.pdf [Accessed 22 October 2017]

Kohonen, T., 1990. The Self-Organizing Map. *Proceedings of the IEEE,* 78(9), pp. 1464-1480.

Maaranen, J., 2010. *Determination of the pump drive energy consumption utilizing process measurements,* Lappeenranta: Lappeenranta University of Techonology.

Muralidharan, V. & Sugumaran, V., 2013. Rough set based rule learning and fuzzy classification of wavelet features for fault diagnosis of monoblock centrifugal pump. *Measurement,* 46(9), pp. 3057-3063.

Muralidharan, V., Sugumaran, V. & Indira, V., 2014. Fault diagnosis of monoblock centrifugal pump using SVM. *Engineering Science and Technology, an International Journal,* 17(3), pp. 152-157.

Rauber, A., 1999. *LabelSOM: On the labeling of self-organizing maps.* Washington DC, IEEE.

Raza, J., Liyanage, J. P., Al Atat, H. & Lee, J., 2010. A comparative study of maintenance data classification based on neural networks, logistic regression and support vector machines. *Journal of Quality in Maintenance Engineering,* 16(3), pp. 303-318.

Rodriguez, P. V. J. & Arkkio, A., 2008. Detection of stator winding fault in induction motor using fuzzy logic. *Applied Soft Computing,* 8(2), pp. 1112-1120.

Sakthivel, N., Sugumaran, V. & Babudevasenapati, S., 2010. Vibration based fault diagnosis of monoblock centrifugal pump using decision tree. *Expert Systems with Applications,* 37(6), pp. 4040-4049.

Selak, L., Butala, P. & Sluga, A., 2014. Condition monitoring and fault diagnostics for hydropower plants. *Computers in Industry,* 65(6), pp. 924-936.

Siemens AG, 2014. *Monitoring of Centrifugal Pumps using the "PumpMon" Function Block in SIMATIC PCS 7.* [Online] Available at: http://support.automation.siemens.com/WW/view/en/42460161 [Accessed 22 October 2017]

Srinivasan, A. & Batur, C., 1994. Hopfield/ART-1 neural network-based fault detection and isolation. *IEEE Trans. Neural networks,* 5(6), pp. 890-899.

Vapnik, V., 1995. *The Nature of Statistical Learning Theory.* New York: Springer.