Chair of Petroleum and Geothermal Energy Recovery

Master's Thesis

Application of machine learning algorithms in assessing the technical condition of pipeline transport facilities

Aliia Siraeva

August 2023

MONTANUNIVERSITÄT LEOBEN
www.unileoben.ac.at

AFFIDAVIT

I declare on oath that I wrote this thesis independently, did not use other than the specified sources and aids, and did not otherwise use any unauthorized aids.

I declare that I have read, understood, and complied with the guidelines of the senate of the Montanuniversität Leoben for "Good Scientific Practice".

Furthermore, I declare that the electronic and printed version of the submitted thesis are identical, both, formally and with regard to content.

Date 29.08.2023

A.Siraeva

Signature Author
Aliia Siraeva

Siraeva Aliia
Master Thesis 2023
Petroleum Engineering / Geoenergy Engineering

Application of machine learning algorithms in assessing the technical condition of pipeline transport facilities

Supervisor: Valeev A., PhD
Co-supervisor/Advisor: R. Albishini, Dip.Ing

Chair of Petroleum and Geothermal Energy Recovery

Abstract

Today, the problem of industrial equipment diagnostics is actively discussed by scientists. First of all, timely diagnosis and early prediction of developing defects directly affect the efficiency of enterprises. For this purpose, enterprises actively introduce new developing technologies, methods and systems, aimed at reducing downtime, energy and material losses and significant increase in socio-economic welfare, as traditional approaches to monitoring the technical condition of equipment and diagnostics, based on periodic inspection does not allow to determine reliably accurate information about the technical condition. One of the innovations are predictive diagnostic systems based on machine learning methods. To date, there are already works and experimental studies of such methods, but further research and validation of such models is needed.

The main purpose of the work is to study the application of machine learning methods to assess the technical condition of objects of the main pipeline transport, on this basis it is planned to develop an algorithm based on data taken from the electric drive of oil pumping equipment.A block based on current and voltage sensors and an Arduino microcontroller for digitization of the obtained data were developed to take the electric characteristics data. To assess the possibility of machine learning algorithms in the work were used such machine learning models as Logistic regression models, and Decision Tree Classificator model.

Through experimentation and application of the developed algorithm, it was found that this algorithm detects the presence of defects with an accuracy of more than 80%, which confirms the potential of integrating physical and digital systems of the production environment

Zusammenfassung

Das Problem der Diagnose von Industrieanlagen wird heute von Wissenschaftlern aktiv diskutiert. Vor allem die rechtzeitige Diagnose und die frühzeitige Vorhersage sich entwickelnder Defekte wirken sich unmittelbar auf die Effizienz von Unternehmen aus. Zu diesem Zweck führen die Unternehmen aktiv neue Technologien, Methoden und Systeme ein, die darauf abzielen, Ausfallzeiten, Energie- und Materialverluste zu verringern und den sozioökonomischen Wohlstand erheblich zu steigern, da die traditionellen Ansätze zur Überwachung des technischen Zustands von Anlagen und zur Diagnose, die auf periodischen Inspektionen beruhen, es nicht erlauben, zuverlässig genaue Informationen über den technischen Zustand zu ermitteln. Eine der Innovationen sind prädiktive Diagnosesysteme auf der Grundlage von Methoden des maschinellen Lernens. Bis heute gibt es bereits Arbeiten und experimentelle Studien zu solchen Methoden, aber weitere Forschung und Validierung solcher Modelle ist erforderlich.

In dieser Arbeit wird die Anwendung von Algorithmen des maschinellen Lernens zur Bewertung des technischen Zustands von Pipelinetransportobjekten untersucht. Das Hauptziel ist die Entwicklung einer Methode zur Diagnose von Ölpumpanlagen anhand der Parameter des Elektromotors unter Verwendung von Methoden des maschinellen Lernens. Zur Erfassung der elektrischen Kenndaten wurde ein Block auf der Grundlage von Strom- und Spannungssensoren sowie ein Arduino-Mikrocontroller zur Digitalisierung der gewonnenen Daten entwickelt. Um die Möglichkeiten von Algorithmen des maschinellen Lernens in der Arbeit zu bewerten, wurden Modelle des maschinellen Lernens wie logistische Regressionsmodelle und Entscheidungsbaum-Klassifizierungsmodelle verwendet.

Als Ergebnis der Experimente und der Anwendung des entwickelten Algorithmus wurde festgestellt, dass dieser Algorithmus das Vorhandensein von Fehlern mit einer Genauigkeit von über 80 % erkennt, was das Potenzial der Integration von physischen und digitalen Systemen der Produktionsumgebung bestätigt.

Table of Contents

Chapter 1

Introduction

1.1 Background and Context

For trouble-free operation of the main pumping equipment (pump-compressor equipment (PCE)) at the pumping and compressor stations, timely diagnostics (or periodic control) of its condition during operational maintenance is of great importance. There is also a strong emphasis on any particular piece of equipment that is most susceptible to wear and tear.

However, the traditional approach to control of technical condition of equipment and diagnostics, based on periodic inspection does not allow to determine reliably accurate information about the technical condition of PCE, which often leads to emergency shutdowns, and also does not allow to effectively plan repair work on the actual technical condition.

1.2 Scope and Objectives

In the midst of the fourth industrial revolution, industrial companies are constantly looking for ways to optimize production lines and reduce costs at the same time. Maintenance costs typically account for more than one-third of total operating costs. Traditional maintenance methods are based on two different strategies: corrective maintenance and preventive maintenance (Susto et al., 2015). The first method means that faulty systems and equipment are repaired when the failure has already occurred, which greatly increases the cost of maintenance. The second is performed at regular intervals to prevent systems/equipment failures. Thus, repairs are performed on machines or components when their service life is not yet known, resulting in unplanned equipment downtime and increased operating costs. Consequently, the most appropriate maintenance strategy should improve the overall condition of the equipment and, in addition, reduce equipment failures, minimize maintenance costs, and maximize equipment life (Carvalho et al., 2019).

1.3 Achievements

In order to improve plant efficiency, reduce downtime, energy and material waste, and significantly improve socio-economic welfare, there is a need to develop a new equipment maintenance strategy. A new strategy scenario would presumably provide better accountability, better cost control and allocation, improved productivity and reduced waste. Under the influence of these factors, predictive maintenance (PdM) has become one of the most urgent topics of the Industry 4.0 era, based on high-quality and optimally planned sustainable maintenance by integrating the physical and digital systems of the production environment. Technological developments in recent years have reduced the costs associated with monitoring the condition of equipment, collecting and storing huge amounts of data. Advanced analytical algorithms provide information on the state of the equipment without interrupting the process and allow strategic decisions to be made on the basis of this information. The main purpose of such monitoring, based on data collection and processing through algorithmization, is to predict a future failure event, which allows for early response and planning of subsequent equipment repairs. In simple words, predictive analytics solves the following problems:

- preventing equipment outages;
- reducing the number of unscheduled repairs;
- evaluation of the technical condition of equipment;
- revealing the causes of equipment failures;
- alerts on approaching failures and malfunctions.

1.4 Technical Issues

Since the volume of data is large enough to be processed continuously by humans or by simple programs that cannot detect the necessary pattern and hence determine the faulty state of technological machines, advances in computing performance have allowed the introduction of machine learning (ML) algorithms capable of finding correlations in complex and large volumes of data. The basic idea behind such algorithms is to collect and train databases in order to create reliable and sufficiently accurate models for equipment diagnostics.

1.5 Overview of Dissertation

Today, machine learning models are being actively implemented in technological processes, including oil and gas operations. There are many different models, but each or their ensemble is selected based on the task at hand. In this paper, to solve the set problem, namely to diagnose equipment by taking the electrical characteristics of the electric drive, considered in the context of the element of pumping equipment, will be considered two of them: logistic regression and decision tree, because the main pump and gas pumping units - powerful energy-intensive machines, the effective operation of which directly affects the reliability and economy of the industry. Along with the problems of technical condition and productivity of the units, it is also necessary to highlight the issue of energy saving during pumping. Failure of responsible equipment leads to additional costs and expenses from downtime.

Chapter 2

Literature Review

2.1 Technical condition monitoring

The history of fault diagnosis and protection is as complex as the machines themselves. Initially, manufacturers adopted simple types of protection, such as overcurrent, overvoltage or ground fault, to ensure reliable and safe operation of electrical machines. However, the advanced functions of these machines required more complex protection systems and more sophisticated systems for diagnostics. Nowadays, it is just as important to diagnose malfunctions at the very beginning, since unplanned machine downtime can disrupt deadlines and lead to large financial losses.

It should be noted that at the moment there is no method, allowing sufficiently reliable and objective identification of defects of oil and gas pumping equipment. Existing methods utilize indirect defect determination based on equipment operating experience. They are also poorly amenable to automation (Valeev A.R., O. M. Saubanov, 2014), (Oscar & Anvar, 2017). In addition, it should be noted that at the moment there are no methods that allow to determine the position of the defect in space, identify the defect and know the degree of its danger, not on the basis of experimental or empirical data, but on the basis of accurate and fundamental analysis.

One of the methods of equipment diagnostics is parametric diagnostics of oil pumping equipment, which is the control of standardized parameters of the equipment, detection and identification of their dangerous changes. When diagnosing gas pumping units, information about power, gas flow rate, efficiency, temperature, etc. is used (Kunina P.S., 2001), (Ponomorev Y.K., 1998) , (Revzin B.S., 2002).

At this type of diagnostics, both analysis of current parameters and analysis of parameters tracking in time - trend analysis are carried out. However, in parametric diagnostics it is impossible to determine the specific cause of the equipment malfunction. The main purpose of parametric diagnostics is to eliminate the consequences of certain changes in equipment operation.

At the boundary of 70-80s of the XX century, mobile and portable vibration measuring equipment, allowing to carry out vibration monitoring of equipment on the basis of frequency analysis, was used in the repair service of productionAt the same time, the problems of equipment reliability theory were actively discussed and solved, which eventually led to the emergence of a new direction - technical diagnostics, the basis of which was the implementation of maintenance and repair (M&R) strategy on the technical condition (TC) (Shirman A.R., 1996).

In the maintenance of rotary machines, vibration monitoring and diagnostics occupy a special place due to the fact that these methods determine the possibility of detecting faults long before the emergency mode is triggered (Barkov A.V., Barkova N.A., Azovtsev A.U., 2012).

According to the works (Barkov A.V., Barkova N.A., Azovtsev A.U., 2012) diagnostics and monitoring are divided into four main constituent parts. The first part of the vibration diagnostics system is the means of measurement and signal analysis. For rotary machines it is the spectrum analysis, based on the primary Fourier transform. The advantage of analysis in frequency bands with relatively constant width is the ability to present a wide frequency range on one graph with a fairly narrow resolution at low frequencies. The resolution in the area of high frequencies (typical both for hydrogasodynamic vibration of different nature, and for gas turbine drives) deteriorates with increasing frequency. In this case, to measure and analyze the spectrum, sensors or digital devices - spectrum analyzers are used.

However, spectral vibrodiagnostics has a number of drawbacks. If the signal level is not stable, the conversion of the signal from one domain area to another will be distorted, resulting in no visual identification of the desired signal peak, instead the graph shows a wide continuous band, which cannot be used to make a correct diagnosis.. Secondly, due to resonances or attenuation of the signal its modulation, both frequency and amplitude, can be observed. In this case, the information fed to the input will be distorted, which also does not provide an opportunity for further research. Third, spectral vibrodiagnostics is characterized by a high level of noise, among which useful signals may be lost, and the defect will be detected when the condition becomes critical. Fourth, spectral vibrodiagnostics requires knowledge of the normal and threshold vibration levels of the equipment to identify the defect at each frequency, in other words, a defect map. This means that spectral vibrodiagnostics is possible only if there is a long experimental operating time of the equipment and an information base on defects (Bogdanov E.A., 2006).

The second part includes a computer with a set of software for vibration monitoring. The main tasks of this software are: storage of measured data, their analysis, comparison of incoming measurements with threshold values.

The third main part of the vibration diagnostics system is the intellectual one, which provides the task of defect identification as well as prognosis of its development in perspective on the basis of measurement data and intermediate materials, prepared by monitoring means in advance.

The last constituent part of the diagnostic system can be considered preventive and timely maintenance means. these means imply performance of general procedures such as balancing of rotors, replacement of bearings, etc.

A machine in perfect working order does not vibrate, but the incomparable presence of imbalances and defects means that even new machines vibrate after repairs. These vibration levels correlate with a serviceable machine, and any deviations characterize the presence of defects or suboptimal operation of the pumping equipment (Barkov A.V., Barkova N.A., Azovtsev A.U., 2012). (GOST 20911-89, 2009.), (Shirman A.R., 1996).

The normative levels of oil trunk pumps vibration intensity are presented in the normative documentation (RD 75.200.00-KTN-119-16 Main Pipeline Transport, 2016). The measurement points are strictly specified (bearing units, casing, supports). Norms for electric motors are presented in (RD 29.020.00-CTN-185-08, 2008). Other norms can be seen in (GOST P 55265.7-2012 [ISO 10816-7:2009]. For gas compressor units the norms of band units of PJSC Gazprom (Temporary standards of bandwidth settings for spectral, 2001) are used. However, this document has a description of only 20 permissible values of the boundaries of areas of RMS values of vibration velocity, as well as the total vibration velocity in the presence of defects in the early stage of development, the possibility of the appearance of secondary defects, the presence of developed defects.

2.2 Experience in diagnostics of oil pumping equipment during its operation in PJSC “Transneft”

Let will consider technical diagnostics of oil pumping equipment in Russia by the example of Transneft, which concentrates a large amount of pumping equipment. Technical diagnostics of oil pumping equipment in PJSC "Transneft" during operation is considered in the context of diagnostic controls with the use of parametric and vibroacoustic criteria. In Transneft, diagnostics is divided into operational, scheduled and unscheduled (RD 153-39TN-008-96). The list of types of control to be performed in Transneft in operation mode is given in (RD 19.100.00-CTN-084-12). Various additional guidelines are given in (RD 23.080.00-KTN-063-11). The main method of inspection during pump operation is vibro-inspection. According to this process documentation, the average quadratic vibration velocity should be in the range of 10 to 1000 Hz. Causes of vibration of pump units in PJSC Transneft are presented in (RD 153-39TN-008-96).

2.3 Technical diagnostics of gas compressor units at PJSC "Gazprom"

At present, the following methods of nondestructive control of the current technical condition are mainly used for diagnostics of gas compressor units (Zaritsky S.P., 1987), (Kunina P.S., 2001):

- vibration diagnostics;
- parametric diagnostics;
- oil condition analysis;

The largest volume of diagnostic information falls on the vibration diagnostics. With its help, about 60% of all defects can be detected, and it is suitable for almost all elements of a gas compressor unit. The remaining defects are usually diagnosed by means of parametric diagnostics and oil condition analysis - respectively, on the order of 20% of the total number of faults (Terentyev A.N., Sedykh Z.S., V.G. Dubinsky, 1979).

2.4 Methods of diagnosing electrical machines

In addition to the method described earlier, there are also a number of others, for example, methods of diagnostics of the electric motor by electric parameters of electromagnetic fields allow to analyze its condition and some general parameters of the whole unit (for example unbalance) quite fully. However, for oil pumping units, attention, as a rule, is shifted more to the pump condition. And, obviously, these methods do not allow diagnosing the whole unit (Petukhov V., 2005).

Methods of diagnostics of electric motors, applied now, demand placing of special gauges directly on engines themselves or it is required full removal of the engine from operation for periodic routine maintenance. Considering this case, we can say that vibrodiagnostics is also applicable, however, it requires special measuring equipment and complex software, which makes this approach quite expensive.

Recently, the EU countries began to actively develop methods of diagnostics of electric machines, based on the monitoring of the state of equipment by the current consumption, followed by the spectral analysis. This analysis allows to determine with a high degree of reliability the technical condition of various elements of motors, as well as with a certain degree of probability to determine the remaining life.

The basic concept of the spectral analysis method summarizes those favorite perturbations in the operation of an electrical machine or part of a mechanical motor led to a change in the magnetic flux in the equipment gap, and as a consequence to a modulation of the current draw.

In (Fedorov D. V., 2007) author also describes the method of applying spectrum analysis by current consumption. The presented technique allows you to determine the following faults and damage:

- inter-winding short circuits of the windings;
- partial or complete breakdown of winding insulation;
- bearing damage;
- rotor imbalance;
- misalignment and run-out of shafts, etc.

The experimental part included direct recording of the current signal, which was carried out during the time required to perform spectral analysis with resolution at a frequency of 0.01-0.02 Hz. The data with the spetcr results are further analysed after being transferred to a personal computer. The diagnosis was established by the presence of characteristic frequencies in the spectrum, which indicate the presence of damage to the electric or mechanical part of the electric motor or an associated mechanical device. For example, a diagnostic sign of motor rotor damage is the presence of two symmetrical peaks relative to the mains frequency (Figure 2.1).

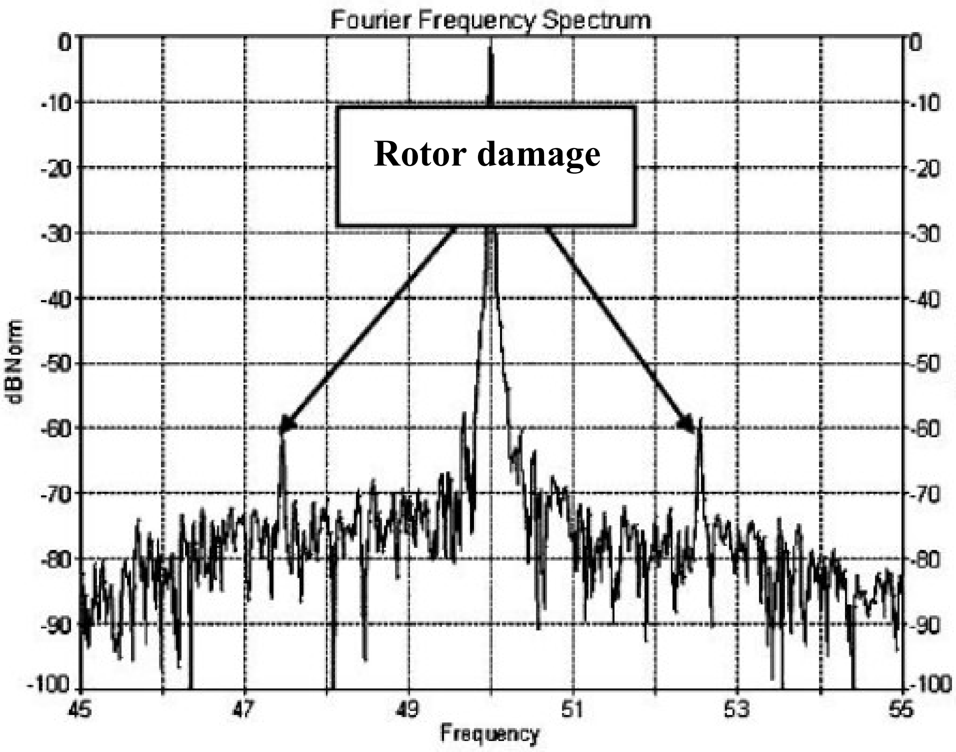

Figure 2.1 - Spectral composition of current of electric motor 5AI-112 M4 with rotor conductor fixing violations

However, the author also suggests monitoring the applied voltage to determine asymmetry or the presence of higher harmonics.

2.5 Wavelet transforms of a signal

Since it is reasonable to carry out equipment diagnostics in the on-line mode, the use of complex expensive equipment that affects the technological process is unacceptable. These requirements, as described above, are fully met by current consumption, or, vibrodiagnostics. Both methods are subject to spectral analysis, which is based on the primary Fourier transform.

The Fourier transform consists in representing the signal *f(x)* as an infinite sum of sinusoids of the form *F(ω)·sin(ω·x)*. The function *F(ω)* - Fourier transform, or spectrum of the signal:

$$F(\omega) = \int_{-\infty}^{+\infty} y(x) \cdot \exp(-i\omega x)\, dx \tag{2.1}$$

The arbitrary signal f(x) can be represented in an orthogonal system:

$$y(x) = C_0\varphi_0(x) + \cdots + C_n\varphi_n(x) + \cdots = \sum_{n=0}^{\infty} C_n\varphi_n(x) \qquad (2.2)$$

where $\varphi_n(x)$ - basis function;

$C_n\varphi_n(x)$ - spectral component $y(x)$;

aggregate coefficients $[C_0, \dots, C_n]$ - signal spectrum. Its graphical representation is shown in a Figure 2.2 called a spectral diagram, which gives a visual representation of the signal spectrum (Yakovlev A.N., 2003), (Astafeva N.M., 1996).

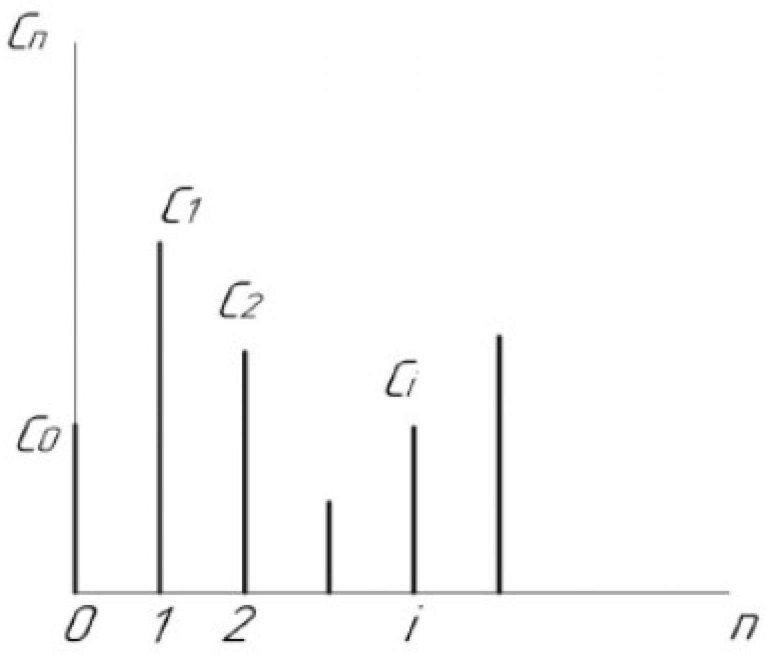

Figure 2.2 - Spectral diagram

[“Despite the many and significant advantages of this method it also has a number of disadvantages, for example, the dynamics of changes in the composition of the signal will be strongly distorted. This disadvantage means that it is not always possible to determine the presence of a defect unambiguously. Moreover, Fourier analysis is extremely unstable with respect to perturbations (shock and dynamic loads). In some cases, this method proves to be ineffective for signals with local features. This is due to the fact that the basis function of the Fourier series is a sinusoid, which is defined in the region from - ∞ to + ∞ and is by its nature a smooth and strictly periodic function. Such a function in practice is fundamentally incapable of describing arbitrary signals and functions. Traditional spectral analysis is ineffective for nonstationary signals with a temporal scale of nonstationarity much smaller than the duration of the realization to be analyzed”] (Kornishin D.V., 2015), (Diakonov V.P., 2002).

These shortcomings do not allow to solve the problem of defect detection comprehensively. The solution to this problem is wavelet transforms of signals, which consider the analyzed time functions in terms of oscillations localized by frequency and time, thus providing a two-dimensional sweep of one-dimensional signals.

The one-dimensional wavelet transform of signal *f(x)* is a two-dimensional function:

$$w_{\Psi f}(a,b) = \frac{1}{\sqrt{C_\Psi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{|a|}} \Psi\left(\frac{b-x}{a}\right) f(x)\, dx \tag{2.3}$$

where Ψ - is the wavelet function;

b - is the shift;

a - is the scale or scale.

The standardizing factor is equal to:

$$C_\Psi = 2\pi \int_{-\infty}^{+\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \tag{2.4}$$

where $\Psi(\omega)$ - the Fourier wavelet image Ψ.

In their essence, wavelet transforms are a representation of a signal in the form of equally shaped short "bursts" that can be shifted and stretched relative to time. This is the fundamental difference from infinite Fourier transform waves.

In (Kruglova T.N., 2019), a wavelet transform method was also applied for more accurate identification of the electric drive state. In this work, the author applied the methodology of recalculating the frequencies of the Fourier spectrum into a wavelet scale, which allows to translate the signal analysis into wavelet space to find new signal characteristics that cannot be detected by the Fourier analysis.

Wavelet analysis is expressed as a parameter dependence:

$$\psi_{ab}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{Ob}\right) \tag{2.5}$$

ψ - primary or maternal function;

ψ_{ab} - analytical function. The designations in the (2.5) have been deciphered earlier. If $(a > 1)$ it is an extension of the basic function, $(a < 1)$ compression.

In integral form, the wavelet analysis looks like this:

$$W_{(a,b)} = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\Psi^{*}\left(\frac{t-b}{a}\right) dt \tag{2.6}$$

where the function *ψ(t)* is the wavelet (analyzing, basis, mother wavelet); the parameter *a* defines the size of the wavelet and is called the scale; the parameter b defines the time localization of the wavelet and is called the shift. The symbol * in formula (*2.6*) denotes complex conjugation (Belyaev A.I. et al., 2015).

2.6 Introduction to Machine Learning in technical condition monitoring

The development of Industry 4.0 technologies opens up new opportunities in all areas of modern industry. Today, technologies such as "digital twin" are used to diagnose the condition of electrical equipment, for the realization of which machine learning methods are used. Such models analyze measurement data of various equipment parameters to identify abnormal data, estimate and predict the development of the equipment condition.

2.6.1 Application of Machine Learning to Diagnose the Technical Condition of Industrial Facilities

With the development of various automated control systems, as well as systems with remote access, there is an increasing trend in the amount of data stored, which reflects information on the technical condition of industrial equipment. such changes provide an opportunity for the development of new methods of defect detection (FDD) (Dai & Gao, 2013), (Corradini et al., 2012). FDD methods are classified into two main groups: model-based methods and signal-based methods. In order to describe the normal operating conditions of asynchronous machines (or IM), assumption principles are used which in turn do not take into account various uncertainties and perturbations. this is done in order to simplify the problem, as such modelling is not practical, and all this together leads to the fact that it is difficult to develop a system or procedure for defect detection (FFD) (Soualhi et al., 2013).

The signal-based method identifies and works with a specific feature, which ultimately provides data on the operating or technical condition of the machine. The methods most used for feature extraction are based on time domain analysis, frequency domain analysis and time frequency analysis. On the other hand, in signal-based methods, a pre-acquired data set from IMs in abnormal conditions is necessary for proper fault diagnosis (Gérard-André Capolino et al., 2011). In this context, MCSA (motor current signature analysis) monitoring strategies use the detection and identification of current signatures that indicate normal or abnormal motor conditions, with the advantages of easy installation of current sensors and no evaluation of motor parameters. However, motor current is affected by many factors such as power supply, static and dynamic load conditions, noise, motor geometry, and fault conditions, so MCSA is not effective for saturation effects or magnetic asymmetry (Concari et al., 2013).

Typical methods for extracting features from current signals use frequency domain algorithms. Traditional methods include: fast Fourier transform (FFT), multiple signal classification (MUSIC) and the estimation of signal parameters by rotational invariant techniques (ESPRIT). These methods are too computationally complex to be implemented online, with the exception of the FFT, so improved versions are discussed in (Kim et al., 2013). Another problem related to the current signature is the rigorous characterization and classification of IMs (Joksimovic et al., 2013). MCSA is also used for fault diagnosis: wavelet decomposition extracts information about the system state from the signal over a wide frequency range. The wavelet components can detect various electrical faults, such as rotor rod breakage, end ring segment, and stator phase loss during operation. Another area of MCSA research concerns the identification of mechanical faults occurring in gearboxes and in the rotor-stator coupling. These faults cause torque fluctuations that affect current signals, as described in (Kia et al., 2009).

In (Zhongming Ye et al., 2003), batch wavelet decomposition is used to extract IM signals with varying torque conditions. The FFT method does not work during startup, while the VT - due to its expansion property and frequency-time representation - picks up signal features under different operating conditions.

In (Jung et al., 2006), an online diagnostic system is proposed. In this paper, MCSA applied to stator current signals is used to detect four classes of motor faults: rotor asymmetry, stator winding fault, bearing failure and abnormal air gap eccentricity. Recently, neural networks (NSs) have been applied to rotor faults of three-phase induction machines: a statistical classification and classification of NSs using the representation of the ambiguity plane is presented in (Boukra et al., 2013). Detection of inter-turn short circuits in the stator and rotor windings of an induction generator with a wound rotor is carried out using DC digital current signals and ANN (Toma et al., 2013).

Also, the probability density function (PDF) algorithm is used for fault detection and diagnosis (FDD). This method is well illustrated in (Giantomassi et al., 2015), the basic idea is that the multivariate normal distribution, calculated from the signals of the current IM, can be regarded as a signature of the state of the IM. Such a signature is robust to external perturbations, but sensitive to faults. The focus is on detecting and diagnosing backlash faults in AC motors, rotors with out-of-tolerance geometries, and cracked rotors. In this approach, motor currents are measured to extract statistical information using probability density function estimation.

The procedure is based on the principal component analysis of the measured signal with signal preprocessing. The characteristics are extracted from the data transformed by principal component analysis (PCA), using a kernel density estimation (KDE) enhanced with a fast Gaussian transform, and a point reduction method.

This procedure is computationally intensive as it significantly reduces the cost of core density estimation. The proposed approach allows models to be defined for each engine condition. The Kullback-Leibler divergence is used as a fault index that allows each fault to be identified. Core density estimation methods have never before been used to diagnose faults in the MCSA scenario. In this context, the authors tested the effectiveness of the KDE-based FDD procedure. The algorithm is completely signal-based, no engine parameters are required. The algorithm was tested in various situations. Tests were performed at different load and voltage levels, as well as at different faults: cracking and out of tolerance rotor and backlash. According to the obtained data, the method is used for its further implementation in production.

2.6.2 Application of the Support Vector Method as a Machine Learning Model

Support vector method (SVM) is a type of supervised machine learning. Can be used for classification and regression analysis. SVM works by finding an optimal hyperplane that separates different classes in the input data. The hyperplane is chosen to maximize the difference between the nearest points of each class. These nearest points, called support vectors, are used to define the decision boundary of the model (Voroncov K.V., 2007).

In the context of equipment diagnostics, SVM can be used to analyze patterns in data collected from equipment in order to identify faults or anomalies. The SVM algorithm is trained on a dataset that includes various points that can be both defective and indicative of normal equipment. The algorithm learns to classify new points based on previous data.

One advantage of this method is that it is able to process multidimensional data, in other words, data collected from multiple sensors to diagnose equipment.

SVM, in turn, is divided into a number of classes: classification, which can be either linear or non-linear, as well as regression and outlier detection.

The method of reference vectors is a popular method for analyzing data from various areas. The method is considered optimal for classification and regression under uncertainty or when classes are not linearly separable. The main task of SVM is to obtain the optimal function f(s) representing the dependence of the output variable on the input variables on the training sample:

$$p_d = f(s) = w^T \varphi(s) + b \tag{2.7}$$

где w – weights;

$\varphi(s)$ - is the function transforming the original feature space into the new one, where the accurate regression is simpler;

b - is the free real-valued element (used for scaling and moving the SVM-generated function).

The main goal is to find the simplest form of the function *f(s)* that best fits the analyzed data. Typical SVM optimization aims to minimize the error function:

$$E_\varepsilon(p, p_d) = \begin{cases} |p - p_d| - \varepsilon \; for \; |p - p_d| \geq \varepsilon \\ 0 \end{cases} \tag{2.8}$$

where ε is the admissible difference between the actual p and the obtained values of the parameters $p_d = f(s)$. Thus, the measurement uncertainty is taken into account.

The SVM is trained using a training dataset *L* (in the form of an array (2.9)) storing examples (symptom vectors) for various simulated states.

$$T = L = \begin{bmatrix} s_1 & p_1 & v_{11} \\ s_2 & p_2 & v_{12} \\ & \vdots & \\ s_n & p_m & v_{nm} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{1k} & p_1 & v_{11} \\ S_{21} & S_{22} & S_{2k} & p_1 & v_{21} \\ & & \cdots & & \\ S_{n1} & S_{n2} & S_{nk} & p_m & v_{nm} \end{bmatrix} \tag{2.9}$$

The total regression error on the training set L is defined as:

$$E = \frac{1}{2} \cdot \sum_{i=1}^{n} E_\varepsilon(p, p_d) \tag{2.10}$$

where n – number of examples in L.

The training method for SVM is quadratic programming (S. Gunn, 1998), which consists in finding optimal weights *w* (2.10) (i.e. minimizing the error) for a selected kernel and its parameters. The procedure is to train L and test T datasets, each containing n - parameters describing the behavior of the system. The data set presented above (2.9) works only for unit fault analysis. Each row stores k -symptoms. They are complemented with information and the actual value of the parameter that is the source of the problem. The presence of a fault is described by two additional columns, containing the identifier of the parameter: r_a (where a = (1,..., m)) and one of its values q, chosen for the experiment (v_{ab}, where b = (1, ..., q)). The other parameters in each example are assumed to be nominal (within acceptable values). This makes the data set (matrix reference) suitable for classification and regression tasks.

In (Bilski, 2014), the author explains the importance of accurate identification of induction motor parameters, which can be used for fault diagnosis, optimization of operation and design of control systems using the reference vector method, presented as a machine learning method. The paper describes the methodology used to apply SVM for parameter identification, including data collection and preprocessing, model training, and validation. The results of a study demonstrating the effectiveness of SVM for parameter identification are presented.

Overall, the paper highlights the potential of SVM as a powerful tool for parameter identification of asynchronous motors, capable of providing accurate results with relatively little data preprocessing.

2.6.2.1 SVM for diagnosing bearing faults in asynchronous motors

Bearings are one of the main parts of electrical machines, ensuring that the rotor rotates smoothly and is mounted at a fixed distance from the stator. About 41% of all faults in electrical machines are due to faulty bearings. Rolling bearings (also known as antifriction bearings) are mainly used in small and medium machines, while friction bearings (thrust liquid-film bearings) are used in large machines. Other common causes of bearing failure include overloading, corrosive environments, mechanical wear and misuse (Önel & Benbouzid, 2008), (Harris T., 2001).

Bearing failures occur due to grease contamination, loss or excess grease, overloading, and corrosive environments. Dirt and other foreign particles that are commonly present in the operating environment contaminate the bearing materials. Abrasive structures of various kinds and degrees of abrasion lead to metal corrosion such as pitting or grinding wear, which increases the wear rate of bearings and raceways considerably. Bearing corrosion is caused by water, acids, degraded lubrication, and even sweat from careless handling during installation. Loss of lubrication causes friction between the rotating parts of the bearing, which increases the effect of heating. This excessive heating leads to destruction of the lubricant and accelerates bearing failure. Therefore, it is necessary to detect bearing failure at an early stage to prevent secondary failure.

The two main methods used to detect and diagnose bearing faults are: vibration measurement, temperature measurement. A detailed review of various bearing fault detection and diagnosis techniques was presented in (Wei Zhou et al., 2007), (Stack et al., 2004), (Tandon et al., 2007). For large-sized induction motors, the vibration measurement method is used as one of the most effective methods, but it is not suitable for small power machines because the method is considered to be complicated and expensive for them. (Valeev A.R., 2020). Therefore, the current signature analysis method can be considered advantageous and universal. The analysis of bearing vibration frequencies in the motor current is much more difficult than their detection in the motor vibration.

To solve the problems associated with spectrum analysis, researchers tend to apply intelligent methods, since they can give more accurate results for ambiguous and "noisy" data. For example, in (Om Prakash et al., 2013), the least-squares support vector method was also applied to diagnose bearing failure by the "parameter" - contamination of the rolling bearing lubricant. In this version, the solution is linear equations instead of quadratic programming.

2.6.3 Clustering

Today, large industrial enterprises use the method of threshold values of any parameter to determine the technical condition of the equipment. It is believed that this approach is quite simple and easy to implement.

For example, at present Transneft uses guidelines from (RD 08.00-60.30.00-CTN-016-1-05, 2005), in which the estimation of vibration state is determined by the threshold values of mean square value of vibration velocity, applied equally to a large class of pumps (RD 75.200.00-KTN-119-16 Main Pipeline Transport, 2016).

With such an approach it is difficult to determine the nucleation of a defect and to recognize in advance irregularities in the operation of the pumping equipment. For example, a pump operating at low load may show little or no underestimated signs of defects, but may still "deteriorate" intensively and soon fail.

In practice, oil pumping units operate in different modes, in which, obviously, have different operating parameters. In this case, all parameter values in all modes will not follow the normal distribution law, and therefore the statistical methods based on it to determine the anomalies will not be applicable.

In the context of oil pumping equipment, application of methods using cluster analysis looks interesting. According to this approach, all the initial data reflecting the parameters of equipment operation are placed as points in the conditional space - n parameters - in n-dimensional space. All points are not uniformly distributed in the dimensionless space. Group or cluster are the points lying closest to each other, and the points not included in these clusters are anomaly points.

The paper (Valeev et al., 2022) shows the practical application of the clustering method. To analyze the application of this method in order to detect abnormal state of equipment or signs of developing defect, synthetic initial data about conditional equipment operation were used, i.e. the data was selected so that their main part reflected the normal operation of equipment, as well as "data with anomalies".

The initial data reflect three operating modes, and have information about pump flow rate, pressure drop across the pump, vibration level, node bearing temperature, power, pressure at the pump inlet, efficiency factor.

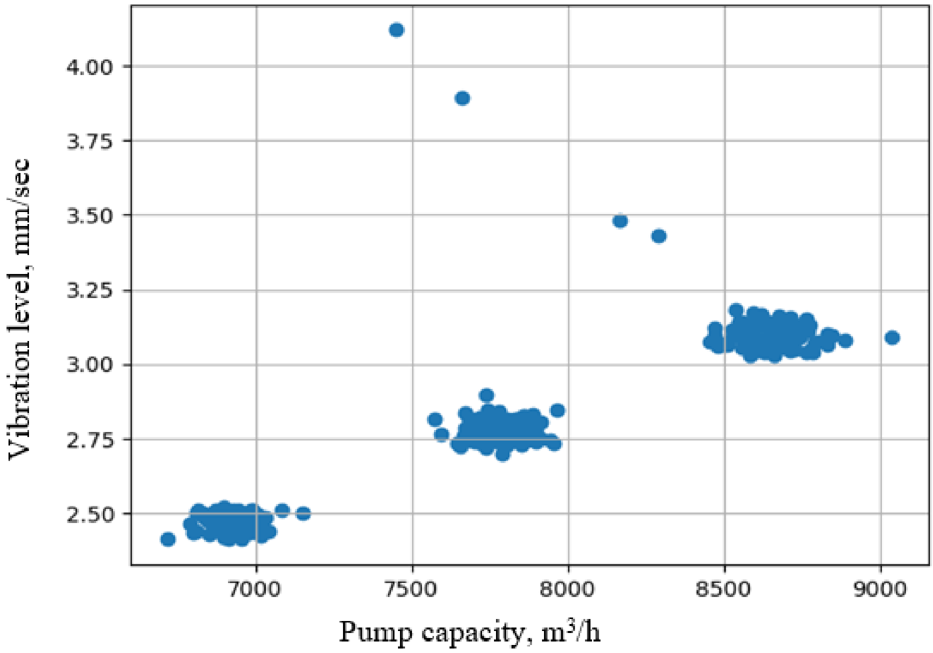

Figure 2.3 – Visualization of the initial data on the pair "pump flow - vibration level" (Valeev et al., 2022)

The number of clusters is taken equal to k, because the k-means method was used. The position of each centroid is defined as the sum of the Euclidean distances from all points to the nearest centroid, and this value must be the smallest. The next step is to plot the dependence of the sum of distances on k, then the optimal number of clusters is determined by the point of the smallest inflection.

Given that the analyzed data had different dimensions, data normalization (in this standard estimation, or z-estimation in another way) was performed. For this purpose, the following formula is used for each i-th dimension of parameter x:

$$z_i = \frac{\overline{x} - x_i}{\sigma} \tag{2.11}$$

where z_i - standardized value of the i-th measurement of parameter x;

x_i- value of the i-th measurement of parameter x;

$\bar{x}$ - average value of measurements of parameter x;

σ - standard deviation of parameter x.

The data are all reduced to dimensionless values with a mean of 0 and a standard deviation of 1.

Statistical analysis determined that a point belonged to the "anomaly" class if at least one value of any of the parameters differed by more than three times the mean standard deviation:

$$s_a = min(\overline{s_i} + 3\sigma_i) \tag{2.12}$$

where $\overline{s_i}$- average value of distances from each dimension to the nearest centroid by i-dimensional parameter;

σ_i - its standard deviation;

s_a - threshold value of distance to the nearest centroid, above which the point is considered anomalous; i - index of enumeration of values by initial parameters.

Therefore, in the original data with information on the operation of equipment identified three normal modes of operation, as well as five abnormal cases.

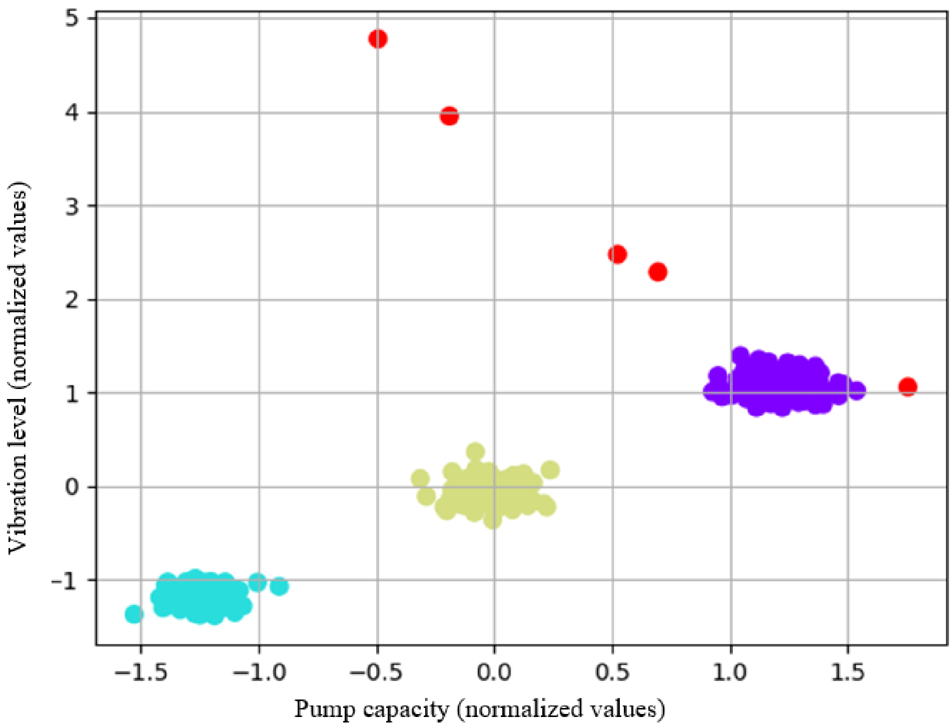

Figure 2.4 – Visualisation of data clustering and highlighting of anomalies (anomalies are marked in red) (Valeev et al., 2022)

To obtain a range of values in which each system parameter is considered normal under different modes of operation, it is necessary to perform a data back-normalization procedure. The results are presented in Table 2.1.

Table 2.1– Range of standard values for equipment operating parameters

Parameter	Regime 1			Regime 2			Regime 3		
	Average	Min	Max	Average	Min	Max	Average	Min	Max
Inflow, m^3/hour	7790	7415	8165	6929	6554	7303	8631	8256	9006
Pressure drop, atm.	15,4	14,0	16,8	13,7	12,3	15,1	17,2	15,8	18,6
Vibration velocity, mm/sec	2,8	2,5	3,1	2,5	2,2	2,7	3,1	2,9	3,4
Temperature, °C	38,9	36,3	41,4	34,6	32,1	37,1	43,2	40,7	45,8
Capacity, kKW	5094	4749	5438	4527	4182	4871	5657	5312	6001
Inlet pressure, atm.	9,3	8,9	9,6	8,2	7,8	8,6	10,3	9,9	10,7
Efficiency	0,802	0,751	0,852	0,711	0,661	0,762	0,885	0,834	0,935

The method of data clustering allows to carry out the procedure of equipment diagnostics both by a separate parameter and by their different combinations.

The calculated range of values obtained as a result of this methodology allows a more accurate assessment of the technical condition of the equipment, also the advantage is that the new threshold values obtained are much smaller than in the established normative documentation. In other words, it allows to detect a defect at its earlier stage, thus reducing the probability of equipment failures, hence providing more reliable transportation of hydrocarbons.

2.6.4 An ensemble of machine learning methods

Machine learning algorithms have been implemented in many technological and everyday processes, and this has also affected the field of equipment diagnostics. Various classification approaches are used, in particular binary classification. These include classical statistical models (naive Bayesian method, classifier, logistic regression, etc.), these methods are focused on machine learning, such as neural networks or reference vector method. The problem is that it is impossible to determine in advance which of the selected methods can solve the problem with the required accuracy, so different methods and their combinations are used, in other words, the aggregate approach.

Obviously, the use of a combined approach would be able to combine the strengths of both approaches within a single model. The study (Nikhil Muralidhar et al., 2018) schematically shows the conditional advantage of using such combined models. In general terms, individual models using only branch diagnostics, as well as models using only machine learning, can produce results with varying reliability. However, models using the combined approach, as shown in the figure below, will achieve the best validity.

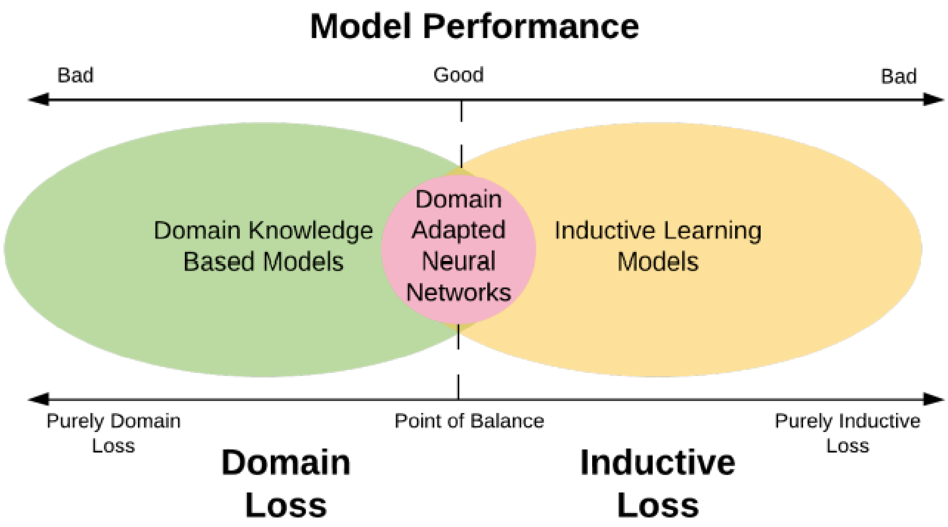

Figure 2.5 - A visual representation of the machine learning model combination approach (Nikhil Muralidhar et al., 2018)

The detailed experiment as well as the theoretical part is well illustrated in the " Forecasting the state of a technical object using machine learning methods " by V. N. Klyachkin. In this paper, the application of machine learning to diagnose the state of a technical object is proposed as a program based on the Statistics and Machine Learning Toolbox library in Matlab, which provides:

-using various basic methods as well as constructing aggregate classifiers;

- application of different criteria for classification;

- changing the volume of the sampling fraction.

The program developed is able to classify an object as serviceable or not by determining the probability at a given threshold value. The program is also able to decide which method or combination to use to solve the problem (Klyachkin & Zhukov D.A., 2019).

2.6.5 Application of Machine learning in practice

Recent research, and its intensity, suggests that machine learning has attracted the attention of both academics and practitioners in industrial service management, emerging as a powerful tool for developing intelligent predictive algorithms in many applications. ML approaches are capable of finding patterns of interest from process or asset data and transforming raw data into a feature space using mathematical algorithms. Recent development in such technological fields as cyber-physical systems or big data allows to obtain much larger data volumes than before, which allows machine learning models to solve more and more different problems, including the tasks of predicting the technical condition of industrial objects

(Wuest et al., 2016). On the other hand, as reported in (Ruiz-Sarmiento et al., 2020), the implementation of ML methods in real-world enterprises remains quite challenging. The main challenges are:

- Data collection and storage (Ruiz-Sarmiento et al., 2017). A large amount of data is used to train the models. In general, more data leads to more robust models and therefore better results. However, the data should carry essential information for further analysis. Therefore, the collection of relevant data has a strong impact on the performance of the ML algorithm.

- Selection and design of the ML algorithm (Wuest et al., 2016). In order to make the right decision about the technical condition of the equipment, the applied machine learning model must provide information about the object in a fairly short time. Although the basic principles of ML are now available to a wide audience, adequate knowledge and deep technical skills are required.

In order to obtain the most accurate results when using ML methods for anomaly detection, several estimation methods should be applied simultaneously, as mentioned above. Such a task has been posed in Pier Francesco Orrù's paper " Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal». The authors propose a preliminary development of a supervised machine learning algorithm ("Supervised Machine Learning") for fault diagnosis of rotating equipment in the oil and gas industry. The fundamental idea is to develop a very simple and executable ML model for rapid and informed decision making. The data was obtained from a real working centrifugal pump, which operates on the production line of the SARLUX refinery plant (Sarroc, Italy). Eight different sensors were used to create the features: one for flow rate, two for bearing vibration, two for axial displacement and three for motor winding temperature. The use of a real dataset has a strong influence on further decisions about the methods to be used for feature development, data labeling, and machine learning techniques. This paper uses and compares two different algorithms, Support Vector Machine (SVM) and Multilayer Perceptron (MLP). The main goal of this paper is not to find a highly accurate and extremely precise MoM model, but to show how a simple and intuitive MoM algorithm can be used to obtain good prediction results. Recent results have shown that the classification results of the proposed solutions (algorithms) are able to determine quite accurately the state of the equipment. The SVM algorithm shows better accuracy than MLP, MLP shows better classification performance, predicting two out of four failures that occurred in the selected period (Orrù et al., 2020).

Chapter 3

Technical Chapters

3.1 General scope and aims

The aim of the following study is analyzed of application of Machine Learning for technical condition monitoring. For this goal experimental study is provided using laboratory installation, developed electric circuit and the machine learning implementation discussed above.

Of the basis of experiment diagnostics data is collected. On the basis of the data various machine learning algorithms are tested.

3.2 Experimental installation for study of machine learning application for equipment diagnostic

Problems of diagnostics of oil pumping equipment are extremely indicative on the example of pump units widely used in modern production of oil and gas industry. To date, there are many potential applications of machine methods for equipment diagnostics, and their study continues. As a rule, organoleptic method is used for the preliminary maintenance, which allows defining the faulty state of the mechanism by detecting the noise, not typical for the normal state, namely tapping, as well as increase of the operating temperature of the equipment. However, this method is extremely ineffective and works only for determining the pre-emergency state.

Since the machine-learning methods for equipment diagnosis described earlier have already found practical application in various other related fields, the use of such solutions for early equipment diagnosis is proposed. This paper discusses the practical application of machine learning techniques such as decision tree and linear regression on an industrial scale.

To identify anomalies in the work of mainline transport objects it was accepted to study electrical characteristics of electric drive - the most common and important receiver in production is an electric motor, because this element of oil pumping equipment system is an integral part of it. The main consumers of electric power in industrial enterprises are three-phase AC motors. Electric load of an electric motor is determined by the magnitude and nature of mechanical load.

The scheme of the experimental setup for approbation of the proposed method is as follows:

Training stand for vibrodiagnostics PROTON-Stand (Figure 3.1) with the possibility of frequency adjustment, including:

- low-noise electric motor;

- pin coupling with shaft;

- speed controller 0-3000 rpm;

- balancing disks with places for installation of balancing masses;

- standard locations for installation of vibration sensors;

- set of balancing weights;

- single platform (frame);

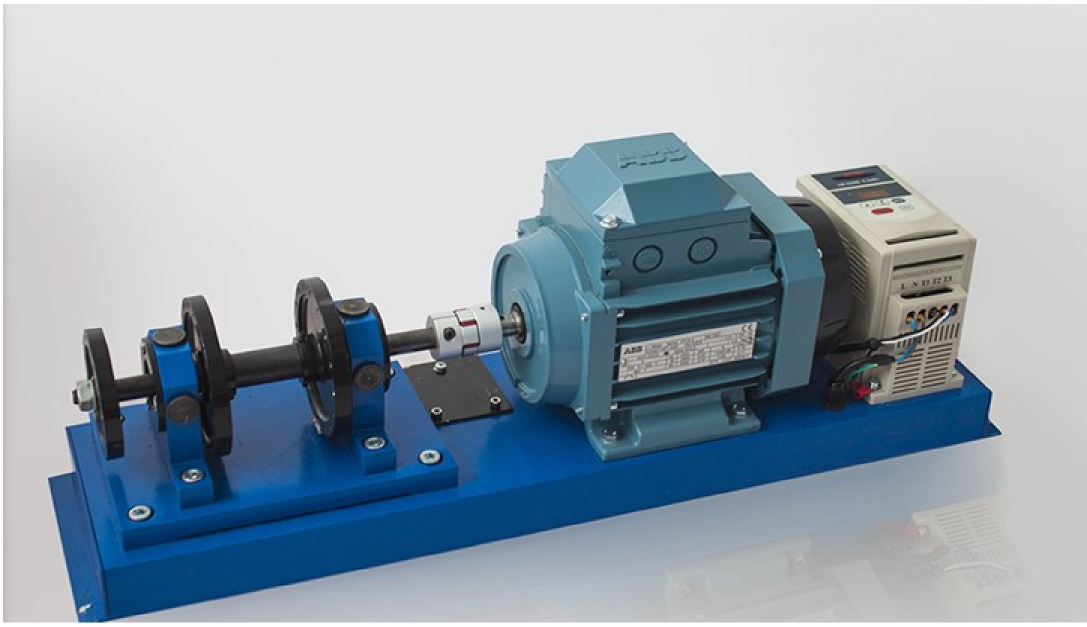

Figure 3.1 - Training model of BALTECH PROTON-Stand centrifugal pump

- A motor electrical characterization block (Figure 3.2), consisting of:

- Arduino Nano controller
- ACS712 30A current sensor
- 1kΩ and 100kΩ resistors

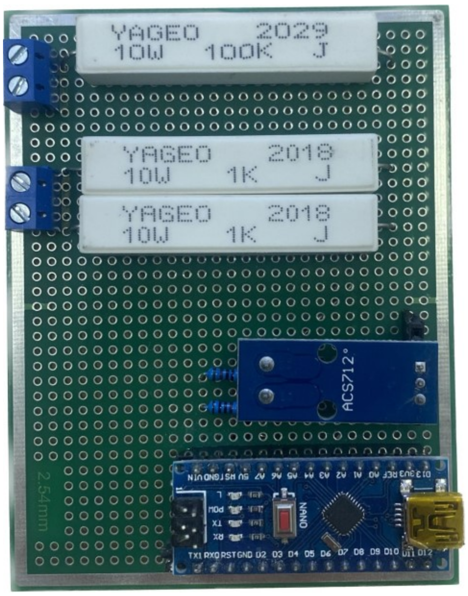

Figure 3.2 - Electrical characterization block

The Arduino Nano platform is a fully functional debug board with its own process and memory. The controller converts the analog signal into digital, then the signal is transmitted to a personal computer by an algorithm written in the Arduino language (see appendix A).

The microcontroller is based on ATmega328 and has the following characteristics:

Microcontroller	**ATmega328**
Operating voltage	5 V
Input voltage (recommended)	7-12 V
Input voltage (limit)	6-20 V
Digital Inputs/Outputs	14 (6 of which can be used as PWM outputs)
Analog inputs	8
DC current through the input/output	40 mAh from one lead and 500 mAh from all leads

Flash memory	32 KB (ATmega328) with 2 KB used for the loader
RAM	2 KB (ATmega328)
EEPROM	1 KB (ATmega328)
Clock rate	16 MHz
Dimensions	1.85 cm x 4.2 cm

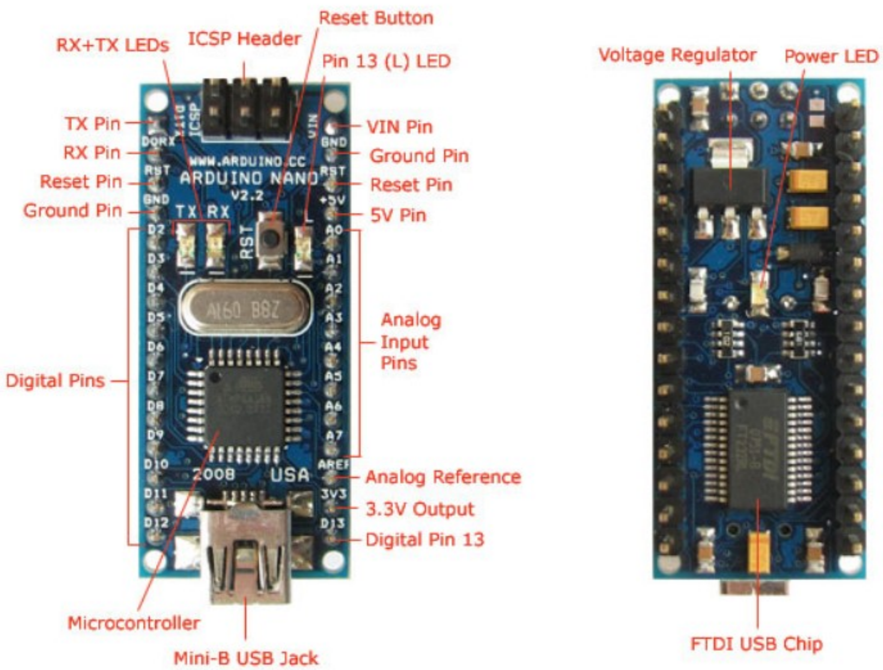

Figure 3.3 - Arduino Nano Pin Description/Pin Diagram

ACS712 30A Current Sensor Specifications:

- Supply voltage +5.0 V;
- Current consumption does not exceed 11mA;
- Current bus resistance 1.2 mΩ;
- Operating temperature -40°C...+85°C;
- Size 31mm x 13mm.

To connect the ACS712 to the Arduino board use 3 wires (Figure 3.4):

VCC - power supply (5V reference voltage);

GND - ground;

OUT - signal (connected to the analog pin of the Arduino controller).

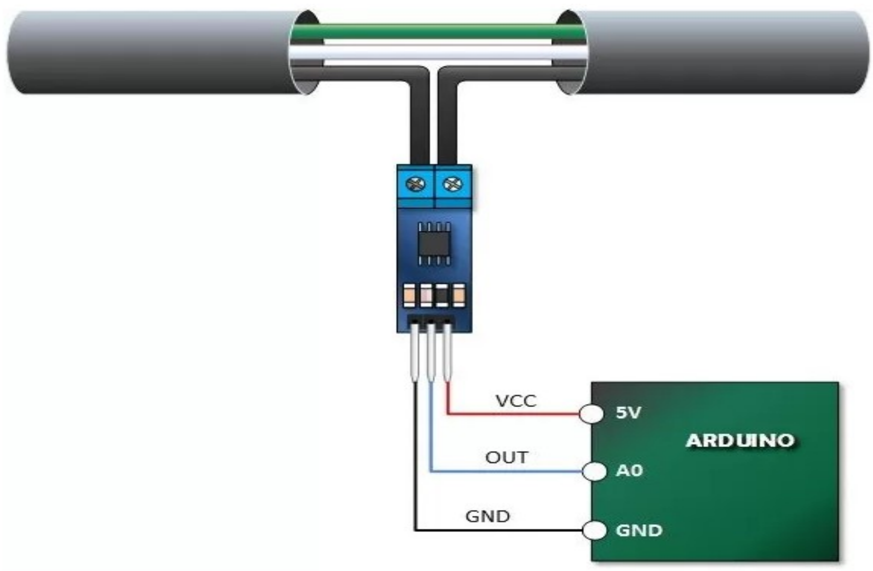

Figure 3.4 - Connecting the ACS712 current sensor to the Arduino board

For measuring the voltage values a voltage divider with resistors was used. The voltage reduced with the divider was brought to values in the range of -2.5V to 2.5V, shifted to a voltage of 0-5V, and fed to the analog input of the Arduino.

The general procedure of data acquisition is as follows - The general procedure of data acquisition is as follows: the microcontroller requests data from 2 channels with a minimum delay of 100 microseconds. The requests are made sequentially, which allows further data acquisition with the maximum frequency of 220,78 Hz. The microcontroller collects a certain amount of data, saves it and sends it to the computer via COM port.

3.2.1 Development of program code

Special programs written in Arduino and Python were developed to receive and analyze the data.

The full code of the program is presented in Appendix B.

The program code conventionally consists of the following parts:

- Sensor block to receive data on the computer. The block receives information via USB cable with additional information about the frequency of the signal measurements from the block.

- The block of primary data processing. In this case the microcontroller acts as an analog-to-digital converter (ADC). Analog signal (measured voltage and current) is fed to the MC, converted to digital. To read the signal the analogRead function is used in the Arduino development environment, after calling this function, the microcontroller measures the level of the analog signal at a given contact, and will save the result of the ADC in a given variable. The result of the analogRead function will be a number from 0 to 1023. Displaying the values, you are looking for in the COM monitor window in the Arduino IDE.

Further signal processing is done using the Python programming language. The obtained data are transmitted through the serial connection using the PySerial library.

To store the obtained data from each experiment, we used a spreadsheet program, Excel. The structure of the tables takes the following form Table 3.1:

Table 3.1 - Example of data storage in Excel

	Current [A]	**Voltage [V]**	**Time [mcs]**
0	-0,4004	0,0250	
1	0,3809	0,0550	
2	-0,4980	0,0750	
3	-0,3125	0,0950	
4	-0,4102	0,0900	
5	1,6113	0,0850	
6	0,9570	0,0900	
7	0,5469	0,0750	
8	-0,4590	0,0400	
9	-0,8887	0,0000	
10	-0,8789	-0,0150	579748,00
11	-0,5664	-0,0550	
12	0,0684	-0,0800	
13	0,0684	-0,1150	
14	-0,0586	-0,1300	
15	-0,0195	-0,1300	
16	-0,6348	-0,1250	
17	-0,7715	-0,1300	
18	-1,6602	-0,1000	
19	-1,5332	-0,7750	
20	0,1270	-0,0500	

To digitize the signal in the time domain, a fast Fourier transform (FFT) of the signal is performed for each experiment with different frequencies. The signal is converted into a spectrum, because the frequency domain is much more convenient for determining the harmonic composition of the signal.

The algorithm of the discrete Fourier transform (DFT), has the following view:

$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi nk}{N}} \tag{3.1}$$

where $x(n)$ is the number of samples of the digital signal

In order to shorten the notation when converting formulas, we replace the variable. Replace the complex frequency $e^{-j\omega}$with the variable W:

$$W_N^{nk} = e^{-j\frac{2\pi nk}{N}} \tag{3.2}$$

The formula for the discrete Fourier transform will now look as follows:

$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)W_N^{nk} \tag{3.3}$$

Write down the sequence of even samples of the signal $x(n)$ in the following form:

$$x_1(n) = x(2n), \qquad n = 0, 1, \dots, \frac{N}{2} - 1 \tag{3.4}$$

A sequence of odd signal samples $x(n)$:

$$x_2(n) = x(2n+1), \quad n = 0, 1, \dots, \frac{N}{2} - 1 \tag{3.5}$$

Now express the discrete Fourier transform through the discrete Fourier transforms of even and odd sequences of input signal samples:

$$X(k) = \frac{1}{N}\sum_{n=0}^{\frac{N}{2}-1} x(2n)\, W_N^{nk} + \frac{1}{N}\sum_{n=0}^{\frac{N}{2}-1} x(2n+1)W_N^{2nk}\ W_N^k\ . \tag{3.6}$$

To measure the scatter of values, we find the mean values as well as the standard deviation in the spectra for two variants: with applied load and under normal operation.

Mean value:

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} \tag{3.7}$$

Standard deviation:

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N}} \tag{3.8}$$

The final part of the developed complex for diagnosing the condition of industrial equipment is the application of machine learning, namely the application of algorithms based on such models as Linear Regression Classificator and Decision Tree Classificator (CART method). The choice of ML models was based on the following criteria:

1) **Simplicity**: Logistic regression models as well as the Decision Tree Classificator model are mathematically less complex than other machine learning methods.
2) **Speed**: Models can process large amounts of data at high speed because they require less computing power, such as memory and processing capacity.
3) **Flexibility**: Logistic regression can be used to find answers to questions that have two or more outcomes. For example, logistic regression can be used to sort data with a large range of values.
4) **Visibility**: The results of the models are easy to visualize using the built-in tools.

3.3 Logistic regression models

Logistic regression is a statistical model that uses a logistic or logit function in mathematics as an equation between x and y. The logit function (sigmoid) maps y as a sigmoid function of x. Constructing a logistic regression equation yields an s-shaped curve (Figure 3.5):

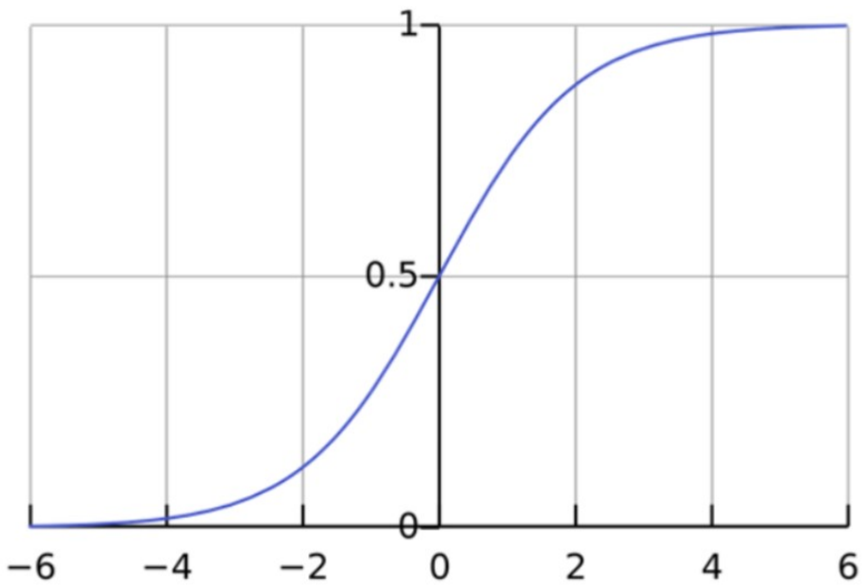

Figure 3.5 - Sigmoid function

The result of a logical regression is the number 0 or 1 for the dependent variable. This is a consequence of this type of regression method, in other words, equations between two independent variables and one dependent variable are solved. This type of regression is excellent for binary classification because, unlike ordinary regression, it assigns the value of the dependent variable to a class, in other words, it determines the probability of assigning the value to a certain class. Consequently, such a dependent variable can have only one value, for example, "0" or "1". The calculated probability value is rounded to the nearest number.

Typically, answers below 0.5 are rounded to 0 and answers above 0.5 are rounded to 1, so the logistic function returns a binary result.

3.4 Decision Tree Classificator model

Decision trees are a logical classification algorithm based on finding conjunctive patterns.

In general, the structure of a decision tree looks as follows (Figure 3.6):

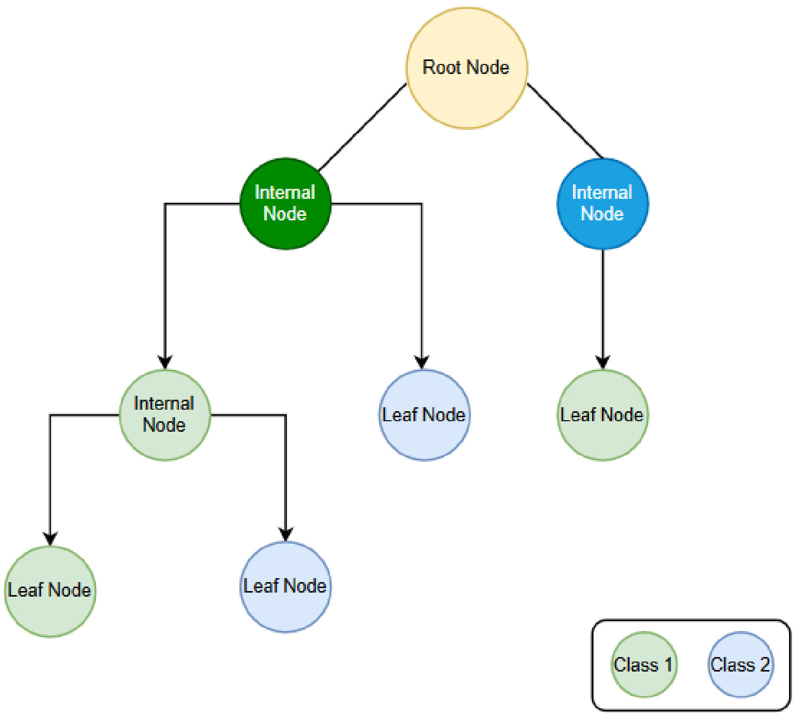

Figure 3.6 - Decision tree structure

- Root Node: Data processing starts in this node.
- Internal Node: This node is also called a decision node because it is the point where data is split into subgroups based on certain criteria.
- Leaf Node: Final node from any internal node, this holds the decision.

The CART method was chosen to build a decision tree model. The decision tree training belongs to the class of learning with a teacher, that is, the training and test samples contain a classified set of examples. The estimation function used by the CART algorithm is based on the intuitive idea of node error reduction. A Gini index is used to analyze the partitioning score of the data, which determines the degree of probability of a particular feature being classified nonparametrically by random selection:

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2, \tag{3.9}$$

where P_i is the probability of assigning an element to a certain class.

3.5 The experimental part

To analyze the possibility of using ML to diagnose industrial equipment the following experimental studies were carried out:

- taking electric characteristics of electric drive under normal operating conditions (without the presence of anomalies and defects);

- reading of electric characteristics of electric drive under conditions of loaded electric drive - a rotor mounted on bearings with high friction was connected;

- artificially created unbalance of the rotor by adding weight (additional mass).

To carry out the experiments, a circuit was assembled from the parts described above, namely a training simulator for vibrodiagnostics PROTON-Stand with the ability to adjust the frequency, a developed unit for the (Figure 3.7):

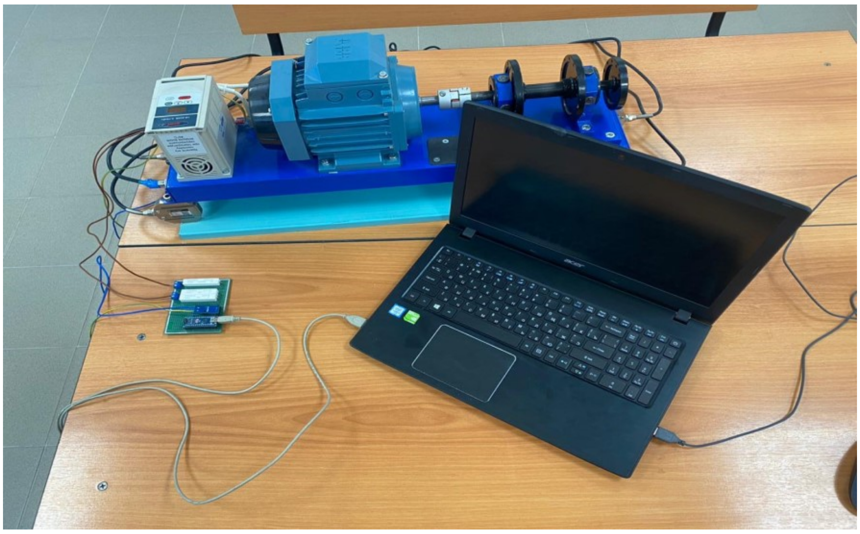

Figure 3.7 - Experimental installation

Since the selected machine learning models are Supervised learning models, a training sample must be fed into the input in order to teach the model to see the relationship between these features and the correct answer. Such a data set in this experiment is the electrical characteristics of a motor taken in the absence of applied load at various frequencies, in other words, in idle mode. The data taken from the motor in this mode are considered reference, since during the initial observation the motor used did not change its temperature characteristics, there were no extraneous noises and vibrations, which indicates that it is in good working order. During the experiment, the variable parameter was the frequency regulated by the variable frequency drive (VFD) (from 20 Hz to 40 Hz in steps of 5 Hz). To evaluate the results the graphs of voltage and current values changes in time were plotted (Figure 3.8):

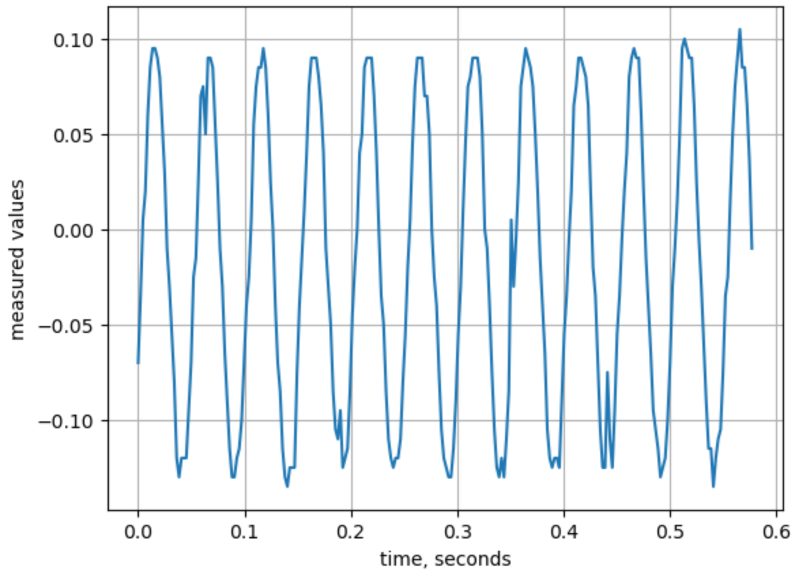

Figure 3.8 - Variation of voltage values in the time domain

Visual assessment assumes that the waveform is a sine wave, which is a periodic alternating voltage (periodic voltage function).

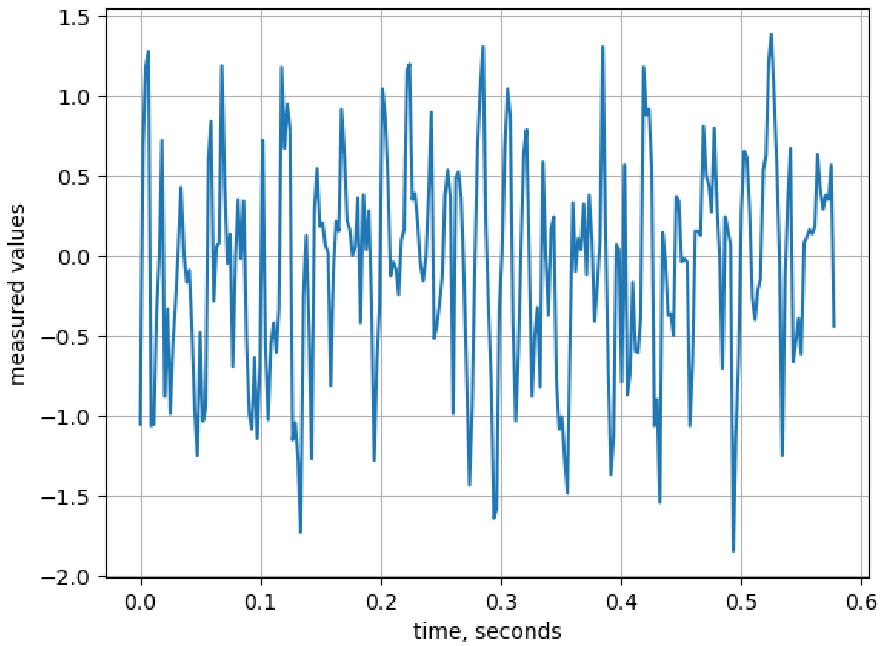

Figure 3.9 - Changing current values in the time domain

Whereas some conclusions can be drawn about the nature of voltage from a graph of voltage variation, the same conclusion cannot be drawn from a current curve, since it represents a more complex law. In other words, these currents have a more complex law of change over time than a simple sinusoidal current. To facilitate the study of this kind of current, it should be known that such a current is a number of simple sinusoidal currents with different amplitudes. Finding harmonics of complex currents is quite difficult. A special section of mathematics, called harmonic analysis, is devoted to this. However, by some signs you can judge the presence of certain harmonics. To simplify the handling of this kind of signal, it is necessary to decompose it into a spectrum.

For a clear assessment on Figure 3.10 - Figure 3.14 presented amplitude-frequency response (AFR) voltage with a loaded rotor considered at frequencies 20, 25, 30, 35 and 40 Hz. Each frequency corresponds to an explicit harmonic amplitude. However, in addition to the fundamental harmonics, some noises are also present, it is especially evident at frequencies 20 Hz and 25 Hz (Figure 3.10 - Figure 3.11). The nature of appearance of these noises can be various, they can be both noises from vibrations, created by operation of electric drive, also the reason can be resonance of auto-oscillation system, another reason of appearance can be so called Flicker noises or noises of 1/f type, this assumption is based on the fact that such kind of noises have distinctive features, for example, noises of 1/f type increase with decreasing frequency, which can be seen on the presented graphs.

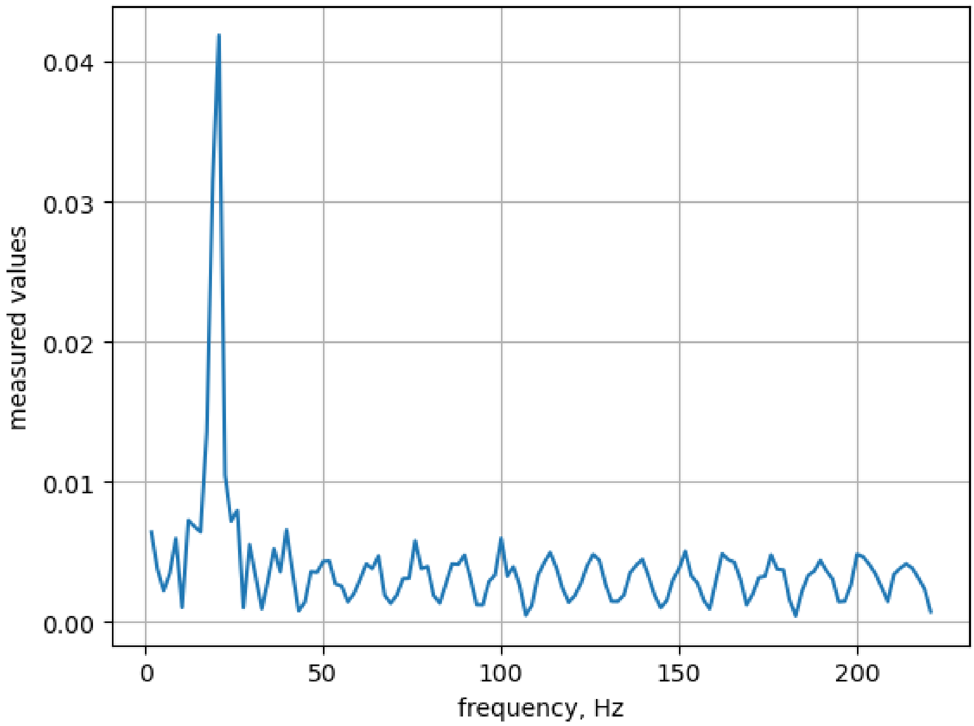

Figure 3.10 - Frequency voltage response at 20 Hz without load

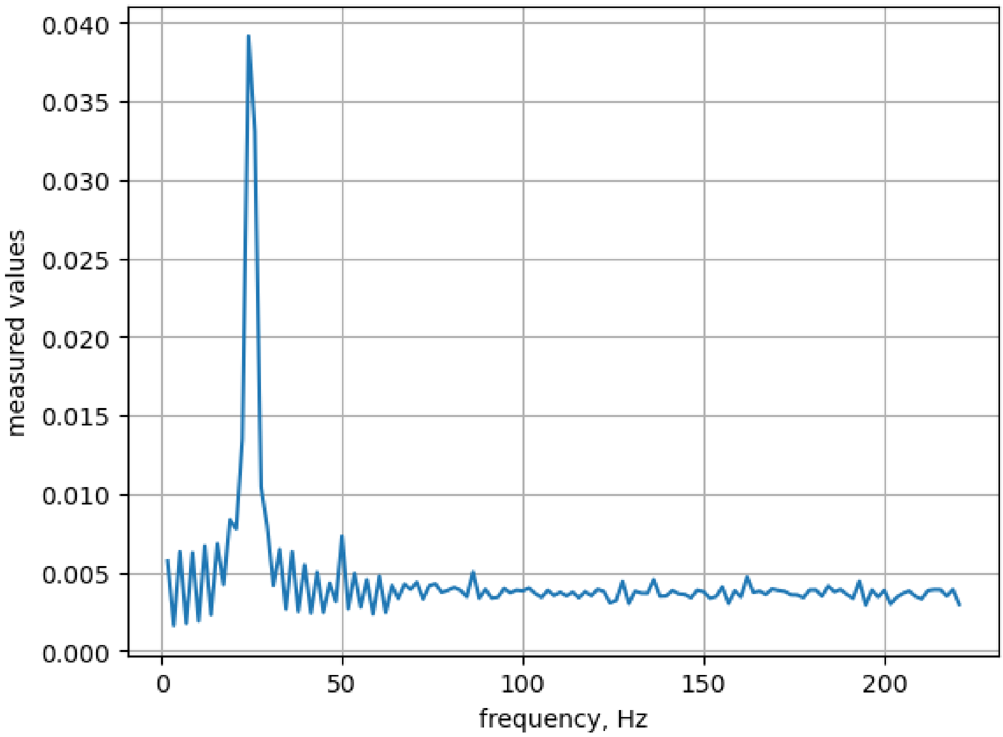

Figure 3.11 - Frequency voltage response at 25 Hz without load

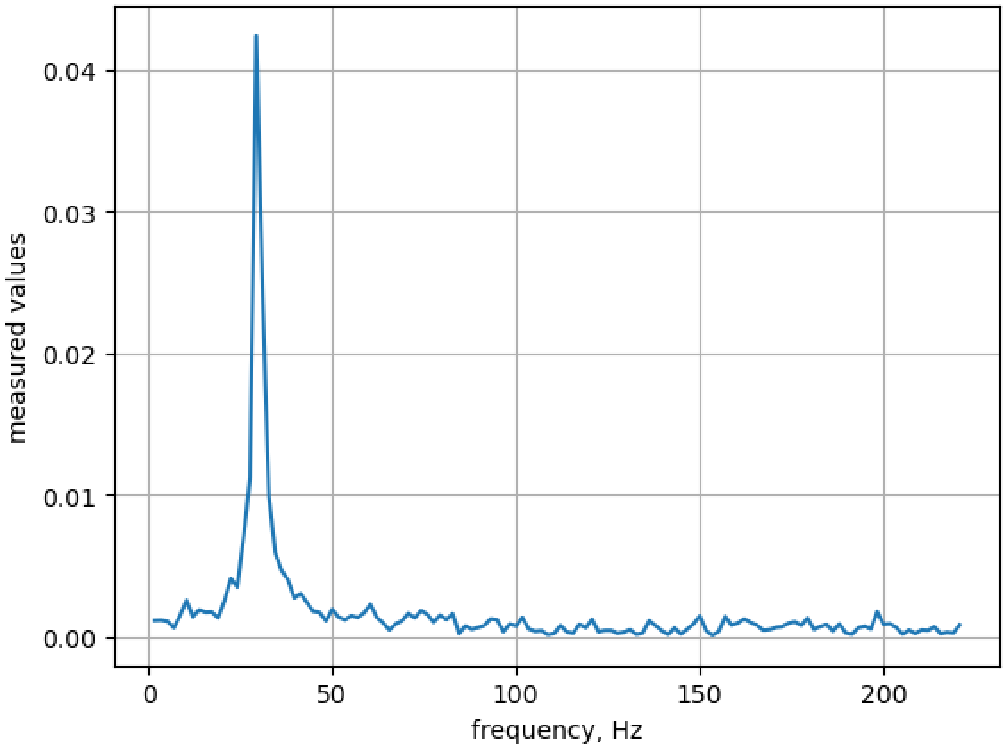

Figure 3.12 - Frequency voltage response at 30 Hz without load

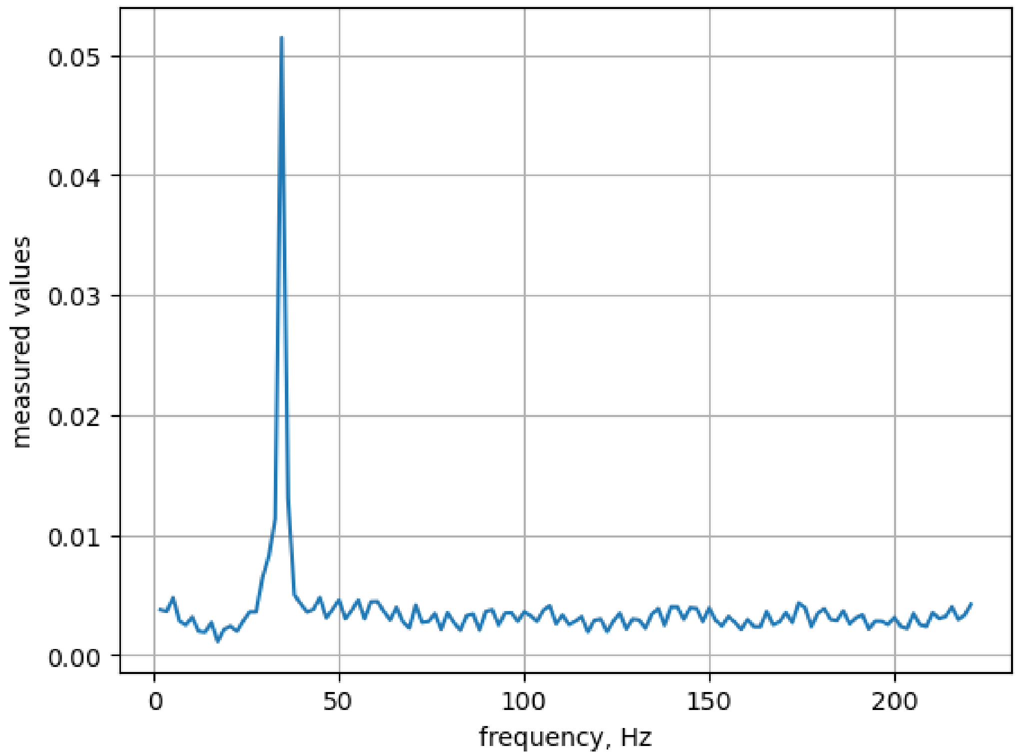

Figure 3.13 - Frequency voltage response at 35 Hz without load

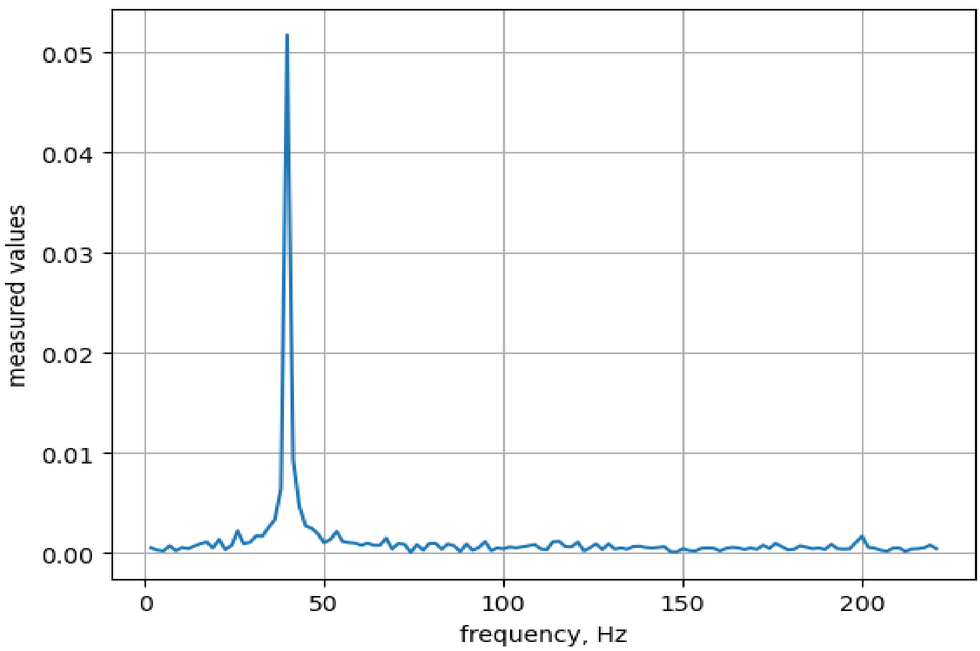

Figure 3.14 - Frequency voltage response at 40 Hz without load

As well as for, voltage on the current-frequency graphs (Figure 3.15 - Figure 3.19), we can determine the peak at frequencies from 20 Hz to 40 Hz in steps of 5 Hz - the fundamental harmonic, but there are others. The nature of the occurrence of these harmonics can be different, most often such "noise" is caused by switching processes. Pulsed power supplies create their own "noise" - usually at frequencies that are multiples of the converter's operating frequency.

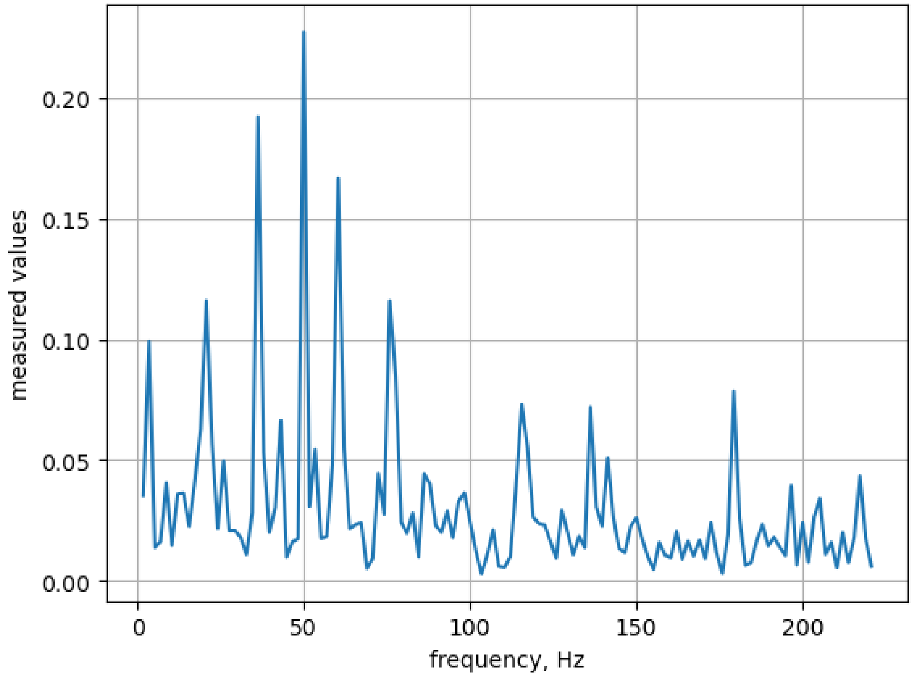

Figure 3.15 - Frequency current response at 20 Hz without load

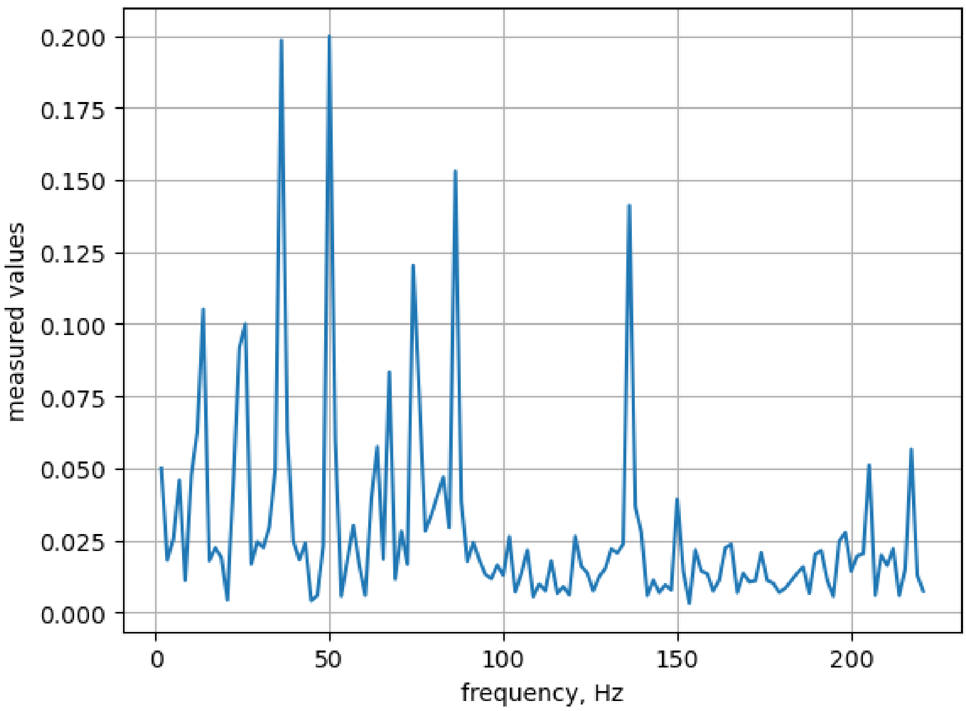

Figure 3.16 - Frequency current response at 25 Hz without load

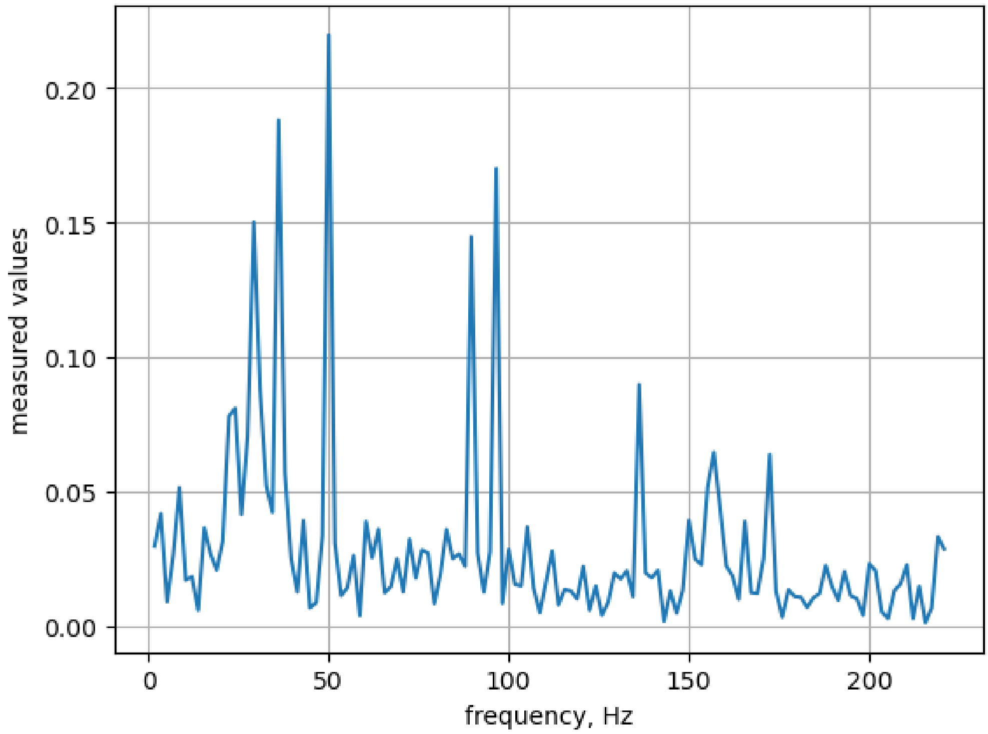

Figure 3.17 - Frequency current response at 30 Hz without load

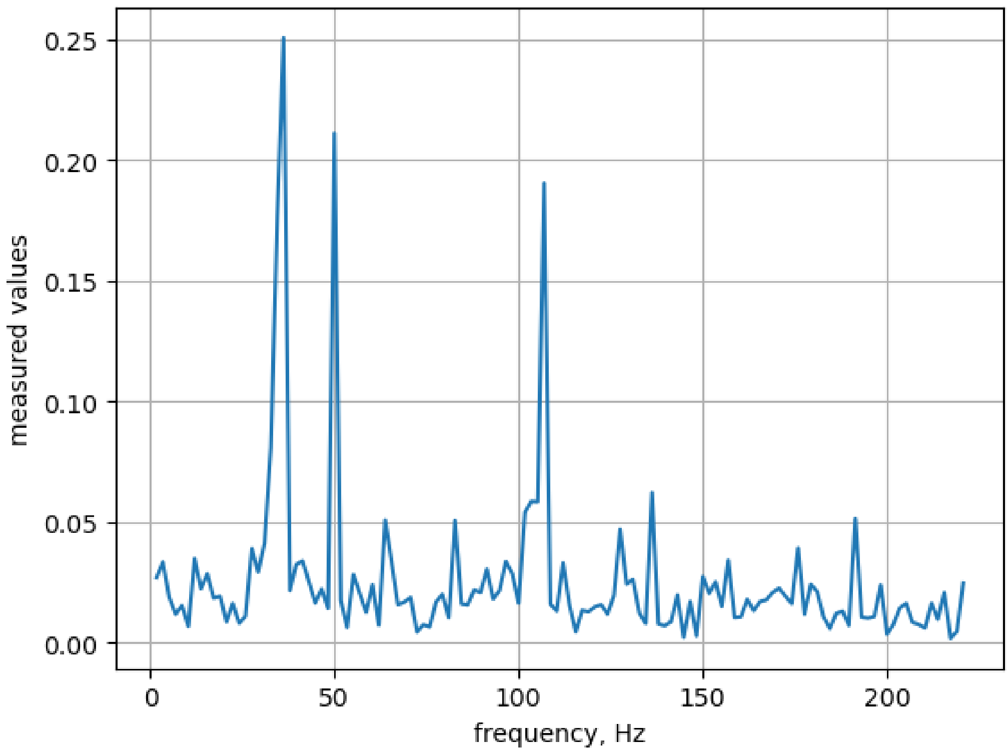

Figure 3.18 - Frequency current response at 35 Hz without load

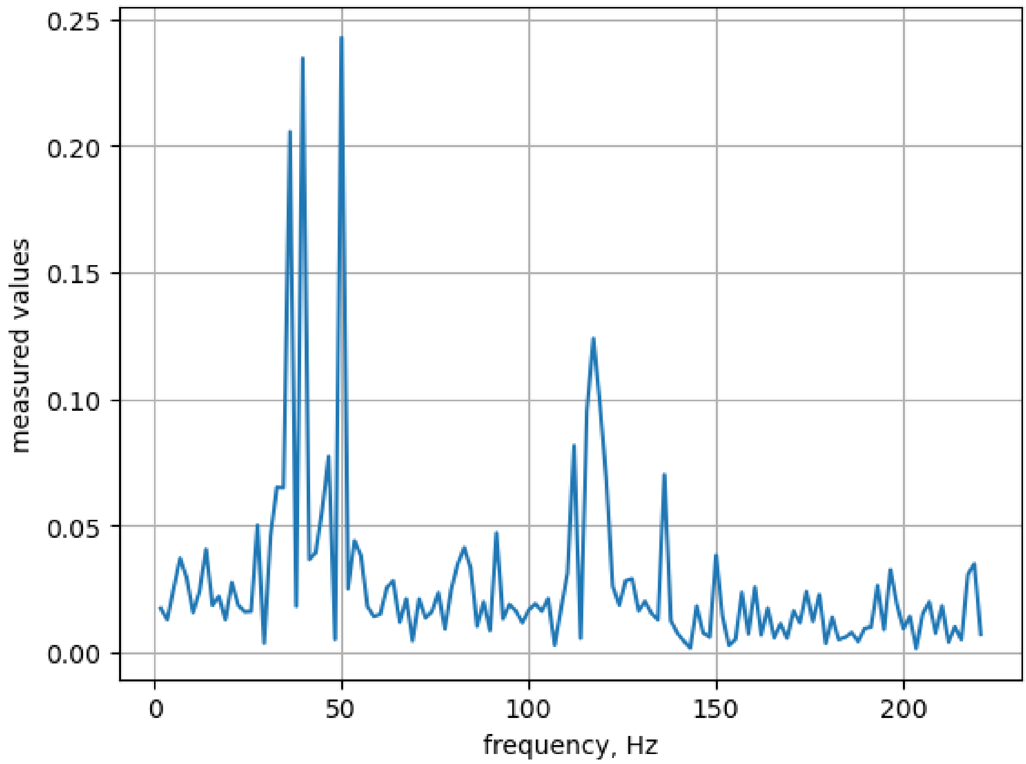

Figure 3.19 - Frequency current response at 40 Hz without load

3.5.1 Rotor unbalance

One of the reliability criteria IM is the uniformity of the air gap. Ununiformity of air gap worsens power characteristics, increases noise and vibration level. The eccentricity of rotor relative to stator axis can be taken as a value of unevenness of air gap. The presence of such eccentricity leads to one-sided magnetic attraction and bearing wear. In this case, periodic contact of the rotor with the stator occurs during IM operation. In the presence of eccentricity, the rotor axis shifts from the set position mainly in the radial direction, causing various pulsations of the electromagnetic field. The amplitude of the rotor oscillation and the frequency of the pulsation depends on the magnitude of the eccentricity, the parameters and the design of the motor.

There are several methods of rotor eccentricity diagnostics: by means of measuring probe; capacitive method; vibration method by means of vibration sensors and method based on spectral analysis of current consumption of stator. The latter two are the preferred methods of fault and defect diagnosis, however, the spectral analysis method is superior in terms of remote monitoring.

As part of the experiment, a weight was attached to the motor shaft to simulate rotor imbalance (Figure 3.20). In this experiment, an artificial change in the air gap between the rotor and stator is created, in other words, a mixed eccentricity is created.

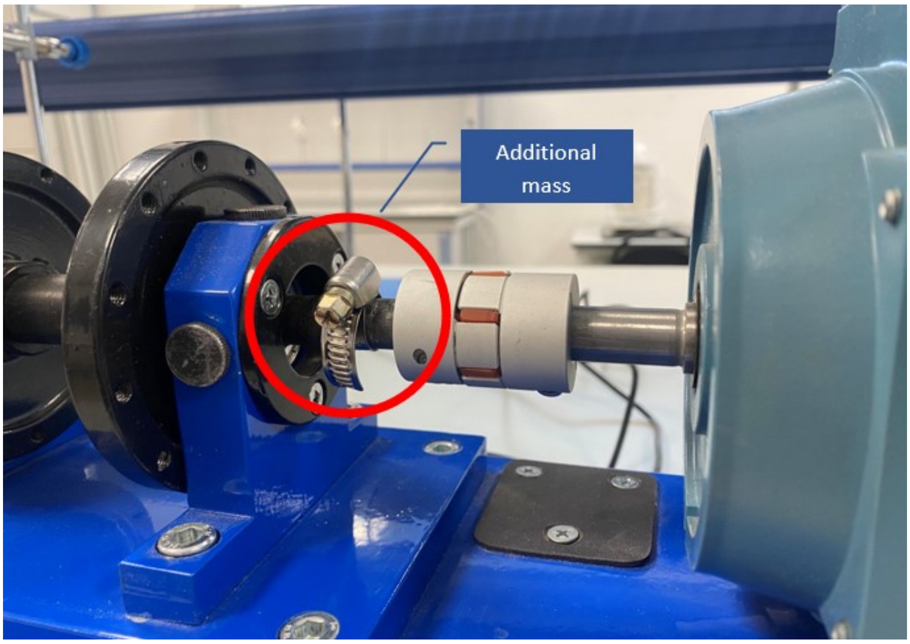

Figure 3.20 - Adding additional mass to the rotor to create rotor imbalance

3.5.2 Comparing of amplitude spectrum characteristics under load, unbalance condition and normal state

The experiments were also performed at five different frequencies, from 20 Hz to 40 Hz in steps of 5 Hz. As a result of the characteristics taken from the actuator, the AFC for each frequency was plotted. A comparative analysis of the spectra at different frequencies was made for both voltage-frequency and current. In the first case, there are characteristic amplitude components, corresponding to the set frequency (Figure 3.21- Figure 3.23).

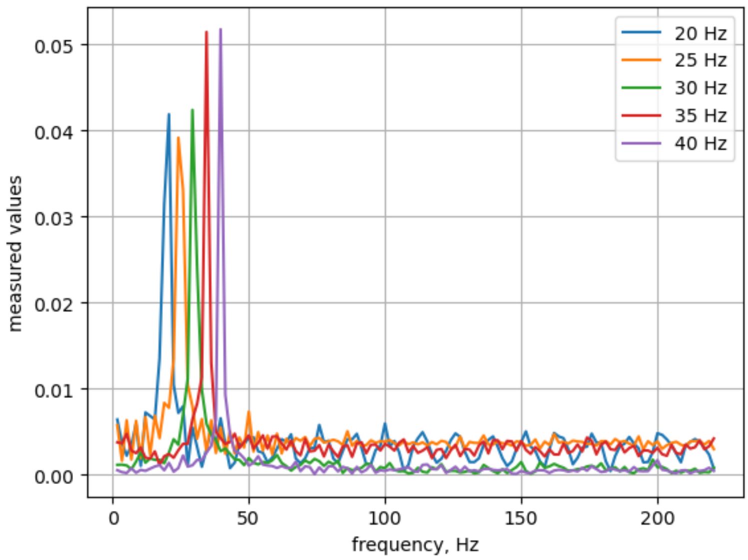

Figure 3.21 - Comparative frequency response of voltage at different frequencies without load

The graph shows 5 spectra of voltages at different frequencies during VFD operation at frequencies 20, 25, 30, 35, 40 Hz. It can be seen that the graph has pronounced peaks at similar frequencies.

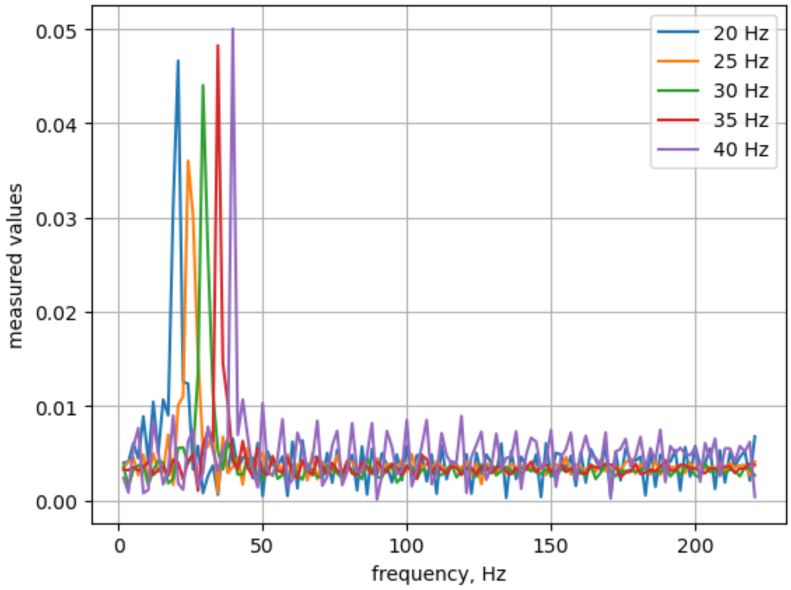

Figure 3.22 - Comparative frequency response of voltages at different frequencies in the presence of a load

The graph shows 5 different spectra at different frequencies from 20 Hz to 40 Hz in 5 Hz steps. It can be seen that the graph has pronounced peaks at similar frequencies. At the same time, the voltage amplitudes differ from the values obtained in the previous graph.

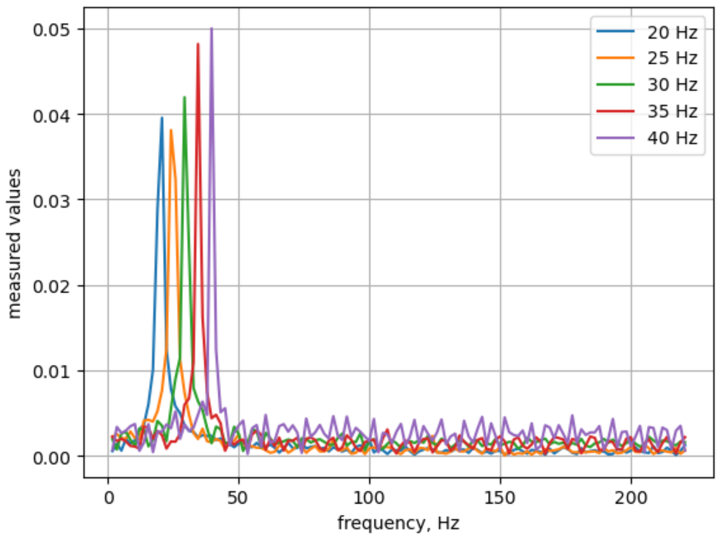

Figure 3.23 - Comparative frequency response of voltage at different frequencies in the presence of load and rotor unbalance

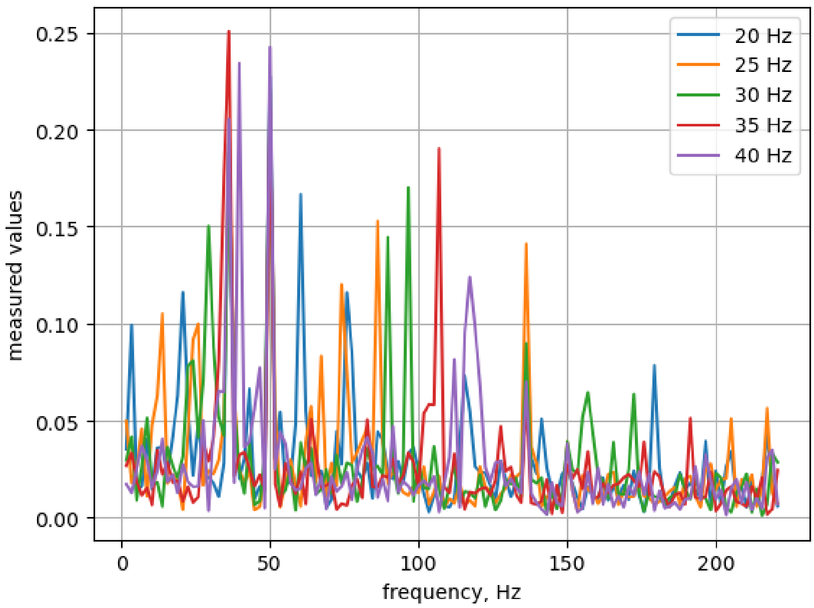

Figure 3.24 - Comparative frequency response of current at different frequencies without load

Analysis of the graphs allows us to conclude that visually it is impossible to notice a specific law of spectrum distribution.

3.6 The average current value

In order to calculate the scatter of the data obtained when the drive is under load compared to the reference data or in other words data taken in idle mode, it was determined that the standard error based on RMS and mean values was 10.2%, which is acceptable for this kind of experiment. In order to clearly evaluate the results, plots of the current dependence on the frequency at 20 Hz were plotted (Figure 3.25, Figure 3.26).

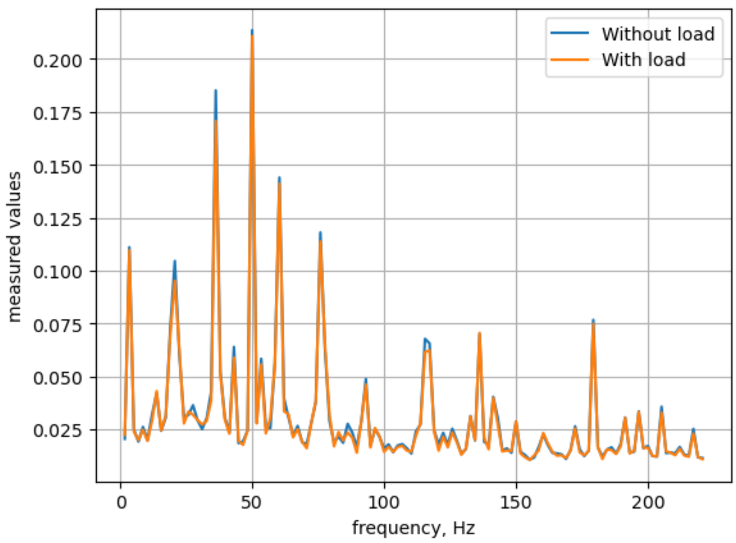

Figure 3.25 - Mean value visualization comparing two spectra with and without load at 20 Hz

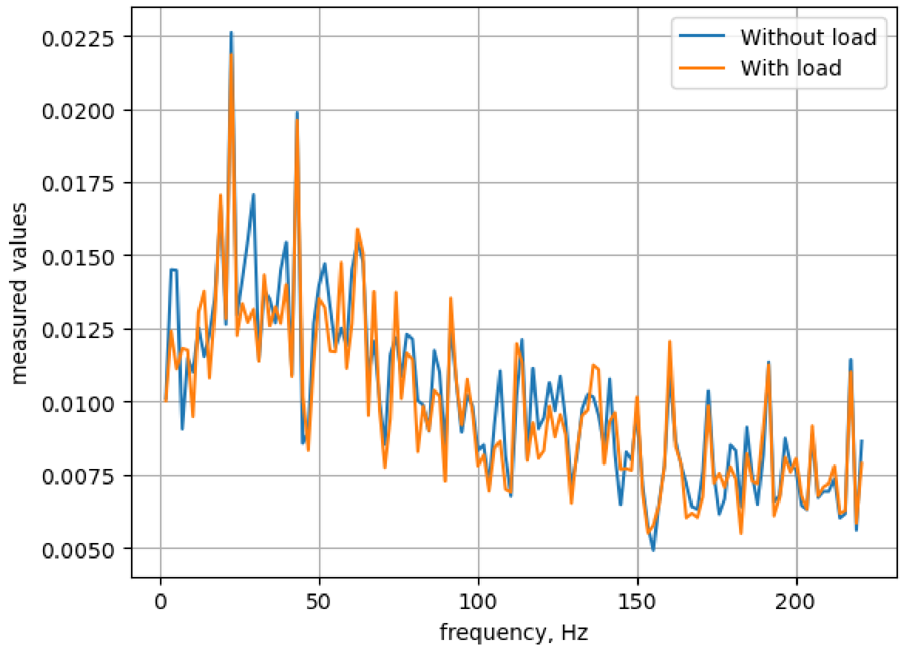

Figure 3.26 - Visualization of the mean square deviation when comparing two spectra with and without loading at 20 Hz

Mathematically, the RMS is defined as the square root of the variance:

$$S = \sqrt{D(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.10}$$

where $\bar{x}$ is the arithmetic mean of the sample:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n) \tag{3.11}$$

x_i-i-th element of the sample;

n -volume of the sample.

3.7 Applying machine learning models

An approach in which in-service diagnostic work enables condition assessment, detection and remedial action during maintenance. For example, the engine condition is diagnosed by fuel consumption, gas temperature, noise and vibration level, composition of exhaust gases, clearance between the cylinder and piston, clearance between crankshaft journals and bearings and other indicators (Birger I.A., 1978). Having numerical values of this set of indicators, it is necessary to assess whether the engine is serviceable or whether it needs to be stopped for maintenance. In this case there is a risk of false alarm (when the serviceable object will be recognized as faulty) or, on the contrary, of missing the target, in which the faulty object is considered serviceable.

In most cases, diagnosis is reduced to the division of object states into two classes: serviceable and faulty. Learning by precedent is one solution to this kind of problem, machine learning is the tool. These methods are reduced to binary classification problems. The known results (precedents) of system state evaluation are considered as input data: the system is working or not working at given values of monitored indicators. Thus, there is a set of situations with given indicators and a set of possible states of the system, which together form an initial sample. This sample is divided into two parts: training and control. The first part is intended for training the model, which in turn classifies the state of the object as serviceable or faulty.

It is assumed that there is some dependence between object functioning indicators and its state. On the basis of this data, it is required to reconstruct this dependence and create an algorithm capable of determining the state of the object with maximum accuracy.

Initial data for diagnostics of the object state, as a rule, are represented in the matrix form - *X* indicators of system functioning, the elements of which x_{ij} is the result of i-th observation on *j*-th indicator, $i = 1, \dots, l$, $j = 1, \dots, p$ (l - number of lines, or number of observations, p - number of columns, or number of indicators) and vector-column *Y*, consisting of zeros for those tests, in which the object is good and units for faulty state. Each row x_i of matrix *X* corresponds to a certain value y_i of vector *Y*. The set of pairs (x_i, y_i) form a sample of initial data - precedents.

The problem is to build a model $a(x, \omega)$ that predicts the answer *Y* for any given *X*. Usually linear models are used for this purpose:

$$a(x, \omega) = \omega_0 + \omega_1 x_1 + \dots + \omega_p x_P, \tag{3.12}$$

$\omega = (\omega_0, \omega_1, \dots, \omega_p)$ - a vector of model parameters (weights).

In this paper, a logistic regression model was chosen as a linear model.

The parameters are selected from the input data; the process of selecting the parameters is called learning the algorithm. The found parameters should provide the optimal value of some quality functional. Often the error functional is minimized (it is the average number of mismatches between the actual result i of the object y_i and the predicted one $a(x_i)$:

$$Q(a,X) = \frac{1}{l}\sum_{i=1}^{l} L(a,x_i) = \frac{1}{l}\sum_{i=1}^{l}\left[a(x_i) - y_i\right] \rightarrow min, \tag{3.13}$$

where $L(a, x_i)$ is the loss function, which captures the presence of a mismatch between the experienced value of the object state for a given set of performance indicators (matrix rows) and the value predicted by the constructed algorithm $a(x_i)$.

Solving this problem, as it was said above, two different methods were used: logistic regression and decision tree. The use of two methods is due to the fact that there is a problem of selecting a method that provides a solution to the problems with the necessary accuracy, so often in practice, combinations of methods are used, the result of which is the ability to decide on the quality functional.

3.7.1 Analysis of the application of logistic regression

Application of the logistic regression model was performed for the whole frequency range: from 20 Hz to 40 Hz in steps of 5 Hz. Both current and voltage data were used for the training and control samples.

As mentioned above, the logistic regression method differs from the ordinary regression method in that in the first case the value of the predicted variable is the assignment of it to a class: "serviceable" or "defective", while in the second case it is the immediate value of the pregnant woman herself that is predicted. To illustrate this, there is some probability belonging to the class of "serviceable" equipment - P_+ , then the probability of the opposite class is defined as $P_- = 1 - P_+$. Thus, the result of a logistic regression always lies in the interval [0, 1].

Visually, the logistic regression can be represented as an n-dimensional space. Since the paper assumes that there are two classes, therefore, there is a two-dimensional space divided by a line into two regions corresponding to the classes (Figure 3.27).

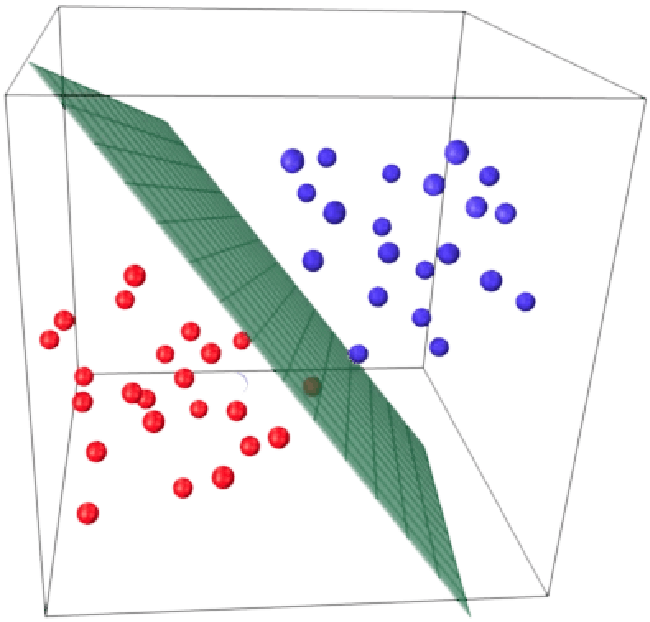

Figure 3.27 - Visual representation of logistic regression in 2-dimensional space

This separating plane is called a linear discriminant, since it is linear in terms of its function, and allows the model to divide, discriminate points into different classes.

The process of dividing variables into classes is as follows: suppose that there are some variables $x_1$and x_2, then the function corresponding to the boundary will take the form:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{3.14}$$

Considering any point, for example (a,b), and substituting the values of $x_1$and x_2, into the boundary function, we obtain:

$$\beta_0 + \beta_1 a + \beta_2 b. \tag{3.15}$$

Since the coordinates of the point change depending on the probability value, it is necessary to consider the following three positions of the point::

1) (a,b) lies in the area bounded by the points of class "+". Then $\beta_0 + \beta_1 a + \beta_2 b$, will be positive, being somewhere within (0,∞). Mathematically speaking, the greater the value of this value, the greater the distance between the point and the boundary. And this means a higher probability that (a,b) belongs to the "+" class. Consequently, P_+ will lie within (0,5, 1].
2) P_+ lies in the region bounded by points of class "-". Now, $\beta_0 + \beta_1 a + \beta_2 b$ will be negative, being within (-∞ ,0). But, as in the case of the positive value, the greater the value of the output value modulo, the more likely it is that (a,b) belongs to the "-" class, and P_+ is in the interval [0, 0.5).

3) (a,b) lies on the boundary itself. In this case, $\beta_0 + \beta_1 a + \beta_2 b = 0$. The model in this case is unable to accurately determine to which class the values a and b belong, so the result P_+ will be equal to 0.5.

For clarity of attribution of values of each measurement an additional column "defect" was added in Excel, because the experimental data are exported to this program.

3.7.2 Analysis of the application of the decision tree model based on the CART method

The binary representation of the decision tree is as follows: in the CART algorithm each decision tree node has two descendants. At each step of the tree construction the rule formed in the node divides the given set of examples (training sample) into two parts - the part where the rule is executed (descendant - right) and the part where the rule is not executed (descendant - left). To select the optimal rule, a partitioning quality function is used - in this context, the Gini index, the algorithm for finding which was described earlier, was applied.

The supposed algorithmic solution is quite obvious. Each node (the general structure of the decision tree is described in chapter 3.4 must have references to two descendants Left and Right - similar structures. Also, a node must contain a rule identifier, describe the right part of the rule in some way, contain information about the number or ratio of examples of each class of the training sample "passed" through the node and have a sign of a terminal node - a leaf. Such a rule in the developed model is a value for voltage or current at different frequency characteristics. In Figure 3.28 - Figure 3.32, the developed model is presented in the form of a hierarchical tree structure. Each tree has been considered for a certain frequency, so the number of measurements is three hundred, where one hundred is the series of measurements referring to the training sample, i.e. assuming fault-free condition, or, in this work, referring to the series of measurements taken at idle running of the electric drive, and, two hundred, referring to the second class - faulty equipment (the measurements taken at artificially created load).

The elements of the decision tree structure are colored in different colors, thus, white indicates uncertainty, blue and its shades indicate defective condition, orange, respectively, belonging to the "defect-free" class.

An influential indicator affecting the accuracy of decision making in the decision tree model is its depth. The larger it is, the lower the error. However, too large a depth value is also not always good, as it can lead to overtraining of the model. As can be seen from the Figure 3.28 - Figure 3.32, the proposed model has a maximum depth of three.

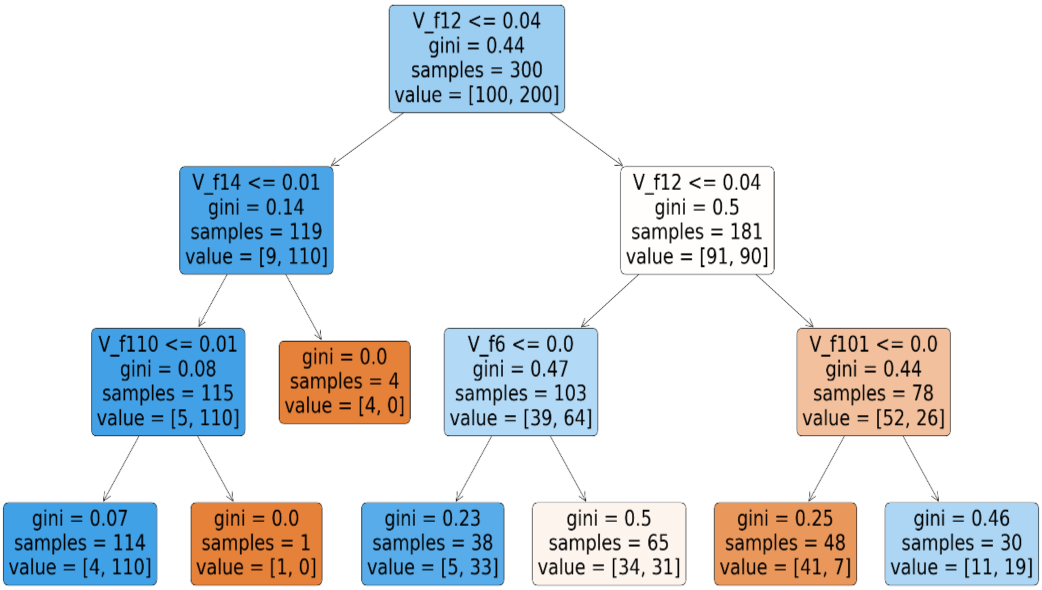

Figure 3.28 - Decision tree for defect detection via voltage values at 20 Hz

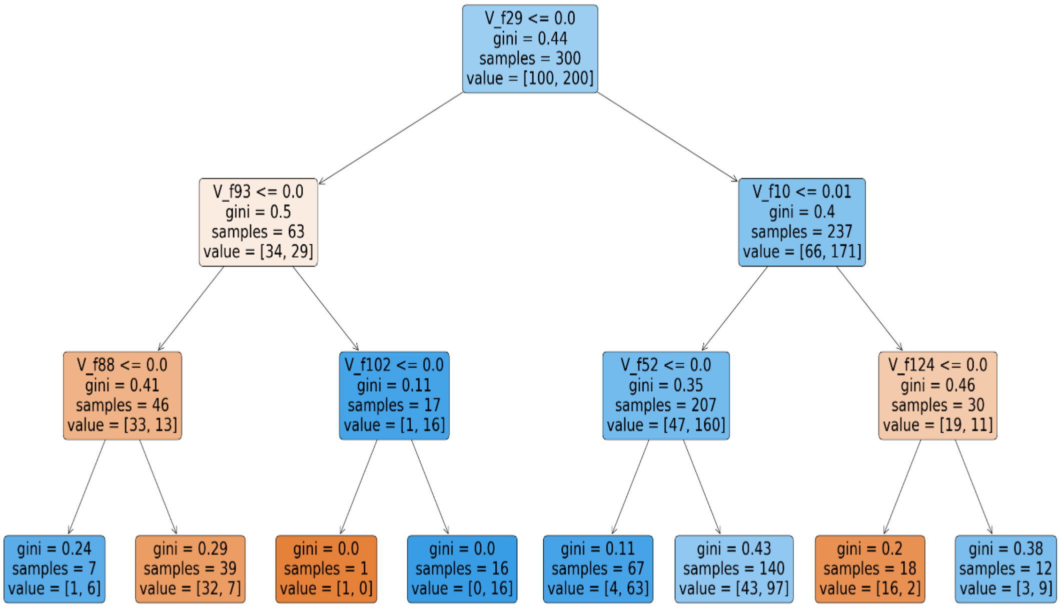

Figure 3.29 - Decision tree for defect detection via voltage values at 25 Hz

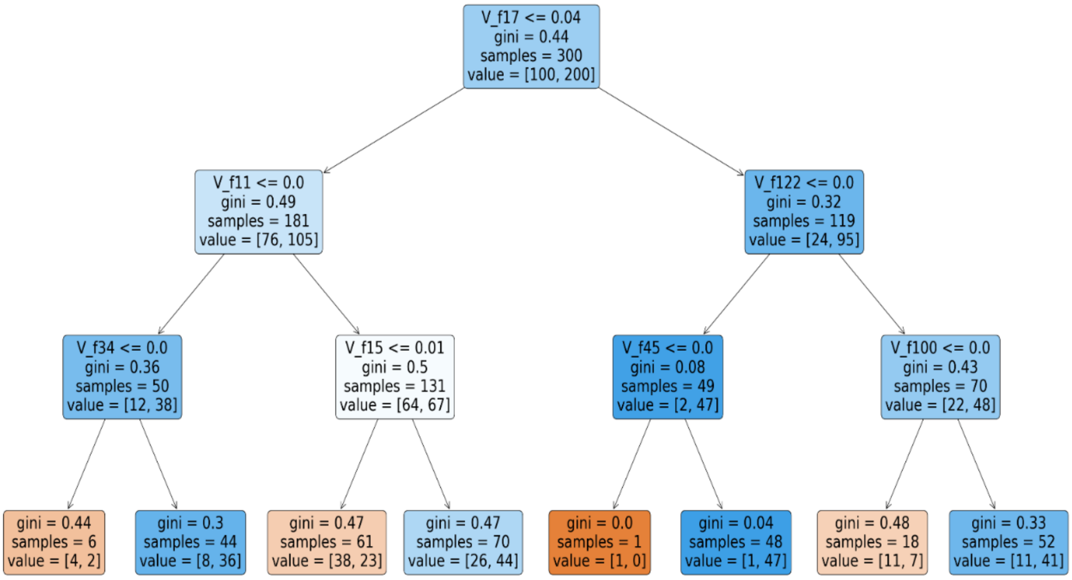

Figure 3.30 - Decision tree for defect detection via voltage values at 30 Hz

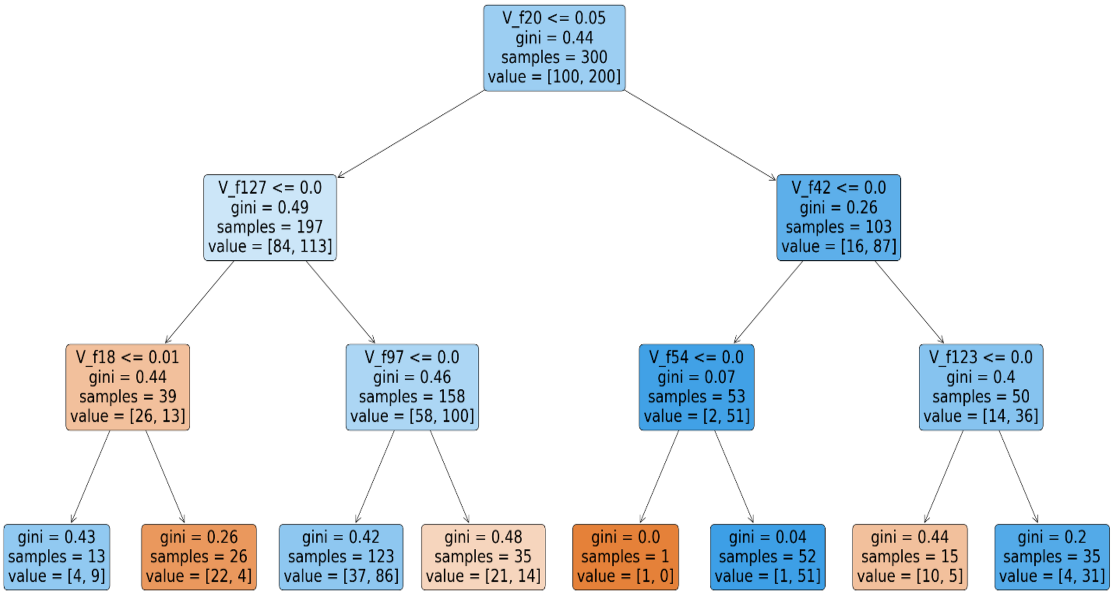

Figure 3.31 - Decision tree for defect detection via voltage values at 35 Hz

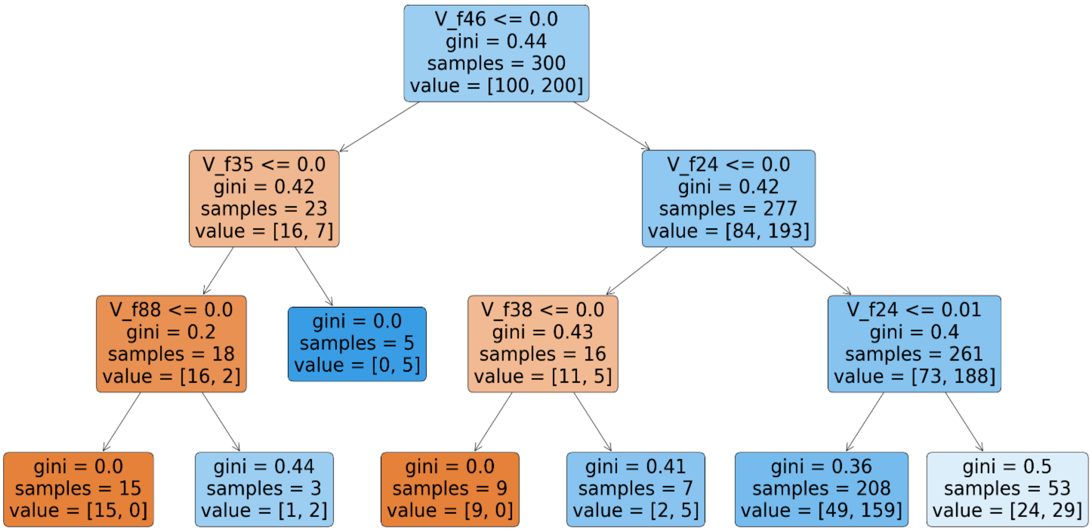

Figure 3.32 - Decision tree for defect detection via voltage values at 40 Hz

Since the initial data were known, i.e. it was known which number of values belong to class "0" and which to class "1", it is possible to determine the accuracy of the model analytically. Thus, only 50.6 values out of 100 were correctly determined as non-defective, which indicates a low accuracy of the model. The same procedure was carried out for the current values, where it was found that only 39.4 out of 100 values were correctly determined. Therefore, it can be concluded that the model should be adjusted more accurately by changing, for example, its depth, hence, to increase the accuracy of this parameter it is necessary to select the most appropriate number of it.

3.8 Model quality assessment based on F-score

In the chapter 3.7 as it was determined before, the F1 metric was used to evaluate the performance of the models, which is the average harmonic value of recall and precision:

$$F = \frac{2PR}{P + R} \tag{3.16}$$

where precision:

$$P = \frac{tp}{tp + fp} \tag{3.17}$$

and recall:

$$R = \frac{tp}{tp + fn}, \tag{3.18}$$

are estimated by the number of correctly classified good states *tp*, the number of incorrectly classified good states *tp* and the number of incorrectly classified states *fn*. The F-criterion, in contrast to the error minimization function, is able to estimate objectively the quality of classification at unbalanced classes (predominance of values of one of the classes over the other).

In turn, the metrics constituting the F1 criteria are based on the error matrix (Figure 3.33). The class that is of interest to us is called "positive", and the remaining class is called "negative". The columns represent the actual values of the target variable, and the rows represent the predicted values of the target variable. For ease of understanding, a decoding of the matrix is presented:

True Positive (TP)

- Assumed value corresponds to the actual value.
- The actual value was positive and the model predicted a positive value.

True Negative (TN)

- Assumed value corresponds to the actual value.
- The actual value was negative and the model predicted a negative value.

False Positive (FP) - Type 1 error

- The predicted value defective state was incorrectly predicted.
- The actual value was negative, but the model predicted a positive value.
- Also known as type 1 error.

False Negative (FN) - Type 2 error

- The predicted value of normal state was predicted incorrectly.
- The actual value was positive, but the model predicted a negative value.
- Also known as type 2 error.

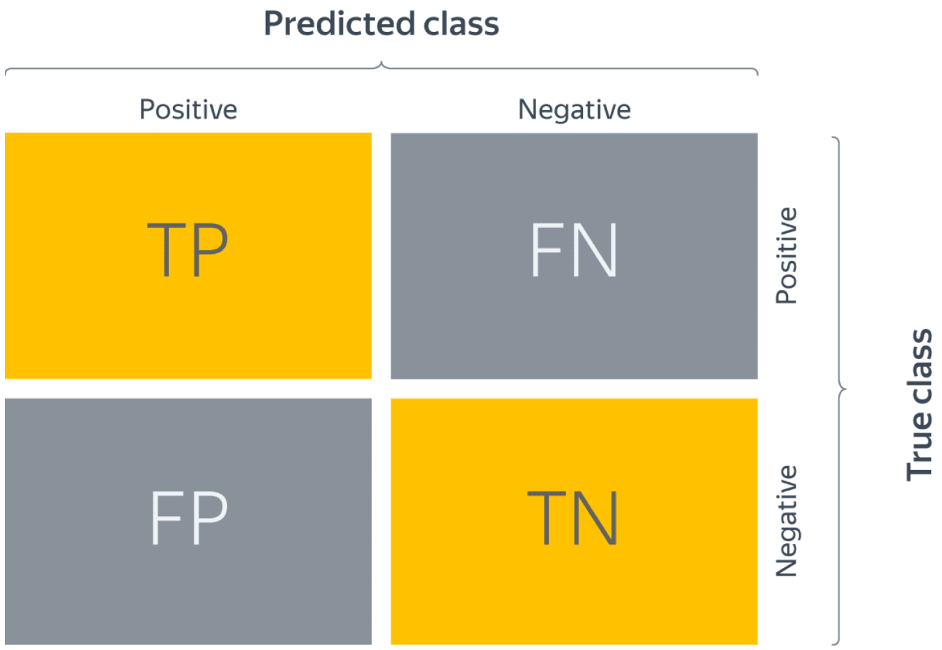

Figure 3.33 - Confusion matrix

F1 metric was used for all proposed models and the whole range of values including all current and voltage values at different frequencies.

The results are presented in the tables:

Table 3.2 - F-1 metric for LogisticRegression and DecisionTreeClassifie models based on current values

Frequency, Hz	LogisticRegression	DecisionTreeClassifier
20	0.680368	0.662967
25	0.70703	0.572197
30	0.48628	0.548564
35	0.387273	0.524199
40	0.396109	0.490219
20, 25, 30, 35, 40	0.143998	0.421023

Table 3.3 - F-1 metric for LogisticRegression and DecisionTreeClassifie models based on voltage values

Frequency, Hz	LogisticRegression	DecisionTreeClassifier
20	0,560162	0,813963
25	0,459445	0,576249
30	0,420972	0,532341
35	0,381662	0,495299
40	0,442377	0,442331
20, 25, 30, 35, 40	0,285791	0,497666

Table 3.4 - F-1 metric for LogisticRegression and DecisionTreeClassifie models based on current and voltage values

Frequency, Hz	LogisticRegression	DecisionTreeClassifier
20	0.681425	0.804247
25	0.69694	0.513071
30	0.476604	0.525286
35	0.39467	0.445239
40	0.396109	0.530988
20, 25, 30, 35, 40	0.127959	0.470489

Evaluating the results of the tables, the following conclusions can be drawn:

1) Application of LogisticRegression and DecisionTreeClassifier models for diagnostics of industrial equipment allows predicting both the priesthood of defects and making conclusion about normal operation of equipment with some accuracy;

2) Comparing the models with each other, we can conclude that the use of LogisticRegression gives poor results for all frequencies, while the predictive ability of the DecisionTreeClassifier model is much higher, which gives reason to believe that this model, the most suitable in the conditions of this work. The relatively poor results of the first model can be explained by the fact that the spectra for different shaft frequencies have different

important bands for analysis, so the decision tree, according to its algorithm, adjusts to different frequencies, while the regression cannot do this.

Chapter 4

Results and Discussion

In the provided study it is offered a method for determination technical condition of industrial equipment via machine learning algorithms. The developed algorithm can be applied to any equipment that has an electric drive. It is based on spectrum analysis of voltage and current of electric drive that measured online. The data obtained from equipment should be transmitted to a computer, where machine learning algorithm of defect prediction is implemented.

On the basis of provided experimental study it is achieved, that the for various frequencies of the equipment it is recommended to use various models:

- for frequency 20 Hz – DecisionTreeClassifier for voltage with metrics of f1-score 0.813963;
- for frequency 25 Hz – LogisticRegression for current with metrics of f1-score 0.70703;
- for frequency 30 Hz – DecisionTreeClassifier for voltage with metrics of f1-score 0.532341;
- for frequency 35 Hz – DecisionTreeClassifier for current with metrics of f1-score 0.524199;
- for frequency 40 Hz – DecisionTreeClassifier for voltage with metrics of f1-score 0.530988;

As it is seen from the results Machine learning works good for small frequencies 20 Hz and 25 Hz. But it works worse for higher frequencies. This suggests that the study of the influence of frequencies on the results requires additional expert and research activities.

It is evident, that LogisticRegression gives no result for whole range of frequencies, but DecisionTreeClassifier deals good with it. This characteristic is easily explained by the fact that the model can adapt to different frequencies under different operating modes. It should also be noted that in order to obtain the most accurate results, it is necessary to adjust the model correctly, i.e., for example, to adjust its depth, but it should also be remembered that this should be done carefully, as there is a tendency to overtrain the model.

It should be mentioned that for better results it is necessary to obtain more precision data, provide more experiments for various defects, analyze more model, it particular more complicated models, such as fully connected neural networks models, autoencoder model, convolutional neural networks.

Chapter 5

Conclusion

5.1 Summary

As a result of the performed scientific research the following tasks were considered:

1) Analysis of modern experience for diagnostics of oil pumping equipment;

2) Design and development of electronic unit for measuring electrical parameters;

3) Conducting experimental research;

4) Processing of experimental data and development of machine learning model.

In order to address the objectives, the following was undertaken:

- Analysis of scientific publications and literature in the field of technical diagnostics;

- Different methods of meshing learning and their application for diagnosing the defective state of equipment were considered;

- In order to analyze the possibility of applying machine learning methods, a clustering method was used.

- In order to test the proposed models, a number of experiments were conducted, for which a block for taking electrical characteristics from the motor was developed.

5.2 Evaluation

The evaluation of the models (Subchapter 3.8) showed that different models are applicable for different operating conditions of the equipment, presumably due to the fact that the VFD affects differently the frequency response at different frequencies. Based on this there is a need to develop the application of different models of a more complex architecture in order to apply them to the operation of equipment at different frequencies.

5.3 Future Work

The limited effectiveness of the model (based on f1-score metric) is due to the simplicity of the model architectures, as well as the use of data with high noise level (low ADT bit rate). For further research, the use of more sophisticated models is recommended, and more in-depth preprocessing of the data is needed. It is also relevant to conduct experiments using different equipment, on a large time interval, the study of a wider range of modes, a larger list of defects and, consequently, the expansion of diagnostic features.

References

Akimov V.I., Saubanov O.M., Kharisov R.M., Valeev A.R., & Tashbulatov R.R. (2018). Разработка комплексного подхода к определению технического состояния насосно-компрессорного оборудования: Development of a comprehensive approach to determining of the technical condition of pumping and compressor equipment.

Astafeva N.M. (1996). Вейвлет-анализ: основы теории и примеры применения.

Barkov A.V., Barkova N.A., Azovtsev A.U. (2012). *МОНИТОРИНГ И ДИАГНОСТИКА РОТОРНЫХ МАШИН ПО ВИБРАЦИИ* (Vol. 156).

Bazhenov V.V. (2006). *Анализ отказов оборудования нефтеперекачивающих станций: Analysis of failures of oil transfer equipment stations.*

Belyaev A.I., A.Y., O., K.A., L., & E.A., D. (2015). Вейвлет-преобразования как метод диагностики энергетических машин. *Инженерный Вестник Дона*, *39*(4-2), 3. https://cyberleninka.ru/article/n/veyvlet-preobrazovaniya-kak-metod-diagnostiki-energeticheskih-mashin

Bilski, P. (2014). Application of Support Vector Machines to the induction motor parameters identification. *Measurement*, *51*, 377–386. https://doi.org/10.1016/j.measurement.2013.12.013

Birger I.A. (1978). *Техническая диагностика: Technical Diagnostics.*

Bogdanov E.A. (2006). *Основы технической диагностики нефтегазового оборудования: Учеб. пособие для вузов - 279 с.*

Boukra, T., Lebaroud, A., & Clerc, G. (2013). Statistical and Neural-Network Approaches for the Classification of Induction Machine Faults Using the Ambiguity Plane Representation. *IEEE Transactions on Industrial Electronics*, *60*(9), 4034–4042. https://doi.org/10.1109/TIE.2012.2216242

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Da Francisco, R. P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, 106024. https://doi.org/10.1016/j.cie.2019.106024

Concari, C., Franceschini, G., Tassoni, C., & Toscani, A. (2013). Validation of a Faulted Rotor Induction Machine Model With an Insightful Geometrical Interpretation of Physical Quantities. *IEEE Transactions on Industrial Electronics*, *60*(9), 4074–4083. https://doi.org/10.1109/TIE.2012.2228142

Corradini, M. L., Fossi, V., Giantomassi, A., Ippoliti, G., Longhi, S., & Orlando, G. (2012). Minimal Resource Allocating Networks for Discrete Time Sliding Mode Control of Robotic Manipulators. *IEEE Transactions on Industrial Informatics*, *8*(4), 733–745. https://doi.org/10.1109/TII.2012.2205395

Dai, X., & Gao, Z. (2013). From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226–2238. https://doi.org/10.1109/TII.2013.2243743

Diakonov V.P. (2002). *Вейвлеты. От теории к практике.*

Fedorov D. V. (2007). Диагностика электрических двигателей на основе анализа спектра потребляемого тока. *Известия Петербургского Университета Путей Сообщения*(2), 69–75. https://cyberleninka.ru/article/n/diagnostika-elektricheskih-dvigateley-na-osnove-analiza-spektra-potreblyaemogo-toka

Gérard-André Capolino, S. Hedayati Kia, & H. Henao (2011). Some Digital Signal Processing Techniques for Induction Machines Diagnosis. In *SDEMPED 2011*. https://www.researchgate.net/publication/271324875_Some_Digital_Signal_Processing_Techniques_for_Induction_Machines_Diagnosis

Giantomassi, A., Ferracuti, F., Iarlori, S., Ippoliti, G., & Longhi, S. (2015). Electric Motor Fault Detection and Diagnosis by Kernel Density Estimation and Kullback–Leibler Divergence Based on Stator Current Measurements. *IEEE Transactions on Industrial Electronics*, *62*(3), 1770–1780. https://doi.org/10.1109/TIE.2014.2370936

GOST 20911-89, 2009. *Техническая диагностика. Термины и определения:* ГОСТ 20911-89.

GOST P 55265.7-2012 (ISO 10816-7:2009). *ГОСТ Р 55265.7-2012 (ИСО 10816-7:2009) Вибрация. Контроль состояния машин по измерениям вибрации на невращающихся частях. Часть 7. Насосы динамические промышленные*.

Gritli, Y., Zarri, L., Rossi, C., Filippetti, F., Capolino, G.-A [Gerard-Andre], & Casadei, D. (2013). Advanced Diagnosis of Electrical Faults in Wound-Rotor Induction Machines. *IEEE Transactions on Industrial Electronics*, *60*(9), 4012–4024. https://doi.org/10.1109/TIE.2012.2236992

Harris T. (2001). *Rolling Bearing Analysis.*

Joksimovic, G. M., Riger, J., Wolbank, T. M., Peric, N., & Vasak, M. (2013). Stator-Current Spectrum Signature of Healthy Cage Rotor Induction Machines. *IEEE Transactions on Industrial Electronics*, *60*(9), 4025–4033. https://doi.org/10.1109/TIE.2012.2236995

Jung, J.-H., Lee, J.-J., & Kwon, B.-H. (2006). Online Diagnosis of Induction Motors Using MCSA. *IEEE Transactions on Industrial Electronics*, *53*(6), 1842–1852. https://doi.org/10.1109/TIE.2006.885131

Kia, S. H., Henao, H., & Capolino, G.-A [G.-A.] (2009). Analytical and Experimental Study of Gearbox Mechanical Effect on the Induction Machine Stator Current Signature. *IEEE Transactions on Industry Applications*, *45*(4), 1405–1415. https://doi.org/10.1109/TIA.2009.2023503

Kim, Y.-H., Youn, Y.-W., Hwang, D.-H., Sun, J.-H., & Kang, D.-S. (2013). High-Resolution Parameter Estimation Method to Identify Broken Rotor Bar Faults in Induction Motors. *IEEE Transactions on Industrial Electronics*, *60*(9), 4103–4117. https://doi.org/10.1109/TIE.2012.2227912

Klyachkin, V. N., & Zhukov D.A. (2019). Прогнозирование состояния технического объекта с применением методов машинного обучения: Forecasting the state of a technical object using machine learning methods. *Международный Журнал Программные Продукты И Системы*, *22*, 244–250. https://doi.org/10.15827/0236-235X.126.244-250

Kornishin D.V. (2015). *Вейвлет-анализ и мультифрактальная параметризация при оценке технического состояния центробежных насосных агрегатов:*.

Kruglova T.N. (2019). Исследование технического состояния электропривода при различных условиях нагружения. *Вестник Белгородского Государственного Технологического Университета Им. В. Г. Шухова*(3), 106–116. https://cyberleninka.ru/article/n/issledovanie-tehnicheskogo-sostoyaniya-elektroprivoda-pri-razlichnyh-usloviyah-nagruzheniya

Kunina P.S. (Ed.). (2001). *Диагностика газоперекачивающих агрегатов с центробежными нагнетателями.*

Kuznetsova M.I. (2015). *Повышение энергоэффективности работы компрессорных станций при эксплуатации газоперекачиювадщих агрегатов с газотурбинным приводом: Improvement of energy efficiency of compressor stations during operation of gas compressor units with Gas turbine drive.*

Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, & Naren Ramakrishnan (2018). Incorporating Prior Domain Knowledge into Deep Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*.

https://www.researchgate.net/publication/330632127_Incorporating_Prior_Domain_Knowledge_into_Deep_Neural_Networks

Om Prakash, Dheeraj Joshi, & G L Pahuja. (2013). *Support Vector Machine based Bearing Fault Detection of Induction Motor* (Vol. 1). https://www.researchgate.net/publication/339301199_Support_Vector_Machine_based_Bearing_Fault_Detection_of_Induction_Motor

Önel, I. Y., & Benbouzid, M. E. H. (2008). Induction Motor Bearing Failure Detection and Diagnosis: Park and Concordia Transform Approaches Comparative Study. *IEEE/ASME Transactions on Mechatronics*, *13*(2), 257–262. https://doi.org/10.1109/TMECH.2008.918535

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., & Arena, S. (2020). Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry. *Sustainability*, *12*(11), 4776. https://doi.org/10.3390/su12114776

Oscar, S., & Anvar, V. (2017). The Monitoring System of an Actual Technical Condition for Pumping Units with Frequency Analysis. *Procedia Engineering*, *176*, 144–149. https://doi.org/10.1016/j.proeng.2017.02.282

Petukhov V. (2005). Диагностика состояния электродвигателей. Метод спектрального анализа потребляемого тока.

Ponomorev Y.K. (1998). *Многослойные демпферы двигателей летательных аппаратов.*

RD 08.00-60.30.00-CTN-016-1-05 (2005). *Руководство по техническому обслуживанию и ремонту оборудования и сооружений нефтеперекачивающих станций.*

RD 153-39TN-008-96. *РД 153-39ТН-008-96 Руководство по организации эксплуатации и технологии технического обслуживания и ремонта оборудования и сооружений нефтеперекачивающих станций*.

RD 19.100.00-CTN-084-12. *РД 19.100.00-КТН-084-12 Правила технического диагностирования и освидетельствования механо-технологического оборудования. Методики технического диагностирования механо-технологического оборудования. Часть 2. Методики выполнения*.

RD 23.080.00-KTN-063-11. *РД 23.080.00-КТН-063-11 Методика определения коэффициента полезного действия магистральных и подпорных насосных агрегатов*.

RD 29.020.00-CTN-185-08, 2008. *РД 29.020.00-КТН-185-08. Положение о системе технического обслуживания и ремонта электротехнического оборудования магистральных нефтепроводов на давление до 10 МПа.* ПАО "Транснефть", 2008.

RD 75.200.00-KTN-119-16 Main Pipeline Transport, 2016. *РД 75.200.00-КТН-119-16 Магистральный трубопроводный транспорт нефти и нефтепродуктов. Техническое обслуживание и ремонт механо- технологического оборудования и сооружений НПС.*

Revzin B.S. (2002). *Газоперекачивающие агрегаты с газотурбинным приводом: Учебное пособие 2-е изд.*

Roslyakov D.A. (2018). *Оптимизация стратегии планово-предупредительных ремонтов магистральных насосных агрегатов: Optimization of preventive maintenance strategy repairs of main pump units.*

Ruiz-Sarmiento, J.-R., Galindo, C., & Gonzalez-Jimenez, J. (2017). Building Multiversal Semantic Maps for Mobile Robot Operation. *Knowledge-Based Systems*, *119*, 257–272. https://doi.org/10.1016/j.knosys.2016.12.016

Ruiz-Sarmiento, J.-R., Monroy, J., Moreno, F.-A., Galindo, C., Bonelo, J.-M., & Gonzalez-Jimenez, J. (2020). A predictive model for the maintenance of industrial machinery in the context of industry 4.0. *Engineering Applications of Artificial Intelligence*, *87*, 103289. https://doi.org/10.1016/j.engappai.2019.103289

S. Gunn (1998). Support Vector Machines for Classification and Regression. https://www.semanticscholar.org/paper/Support-Vector-Machines-for-Classification-and-Gunn/ceb5e9c07f2d95a700c1ed0813dfbae8c3901c18

Shirman A.R., S. A. (1996). *рактическая вибродиагностика и мониторинг состояния механического оборудования.*

Soualhi, A., Clerc, G., & Razik, H. (2013). Detection and Diagnosis of Faults in Induction Motor Using an Improved Artificial Ant Clustering Technique. *IEEE Transactions on Industrial Electronics*, *60*(9), 4053–4062. https://doi.org/10.1109/TIE.2012.2230598

Stack, J. R., Habetler, T. G., & Harley, R. G. (2004). Fault Classification and Fault Signature Production for Rolling Element Bearings in Electric Machines. *IEEE Transactions on Industry Applications*, *40*(3), 735–739. https://doi.org/10.1109/TIA.2004.827454

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812–820. https://doi.org/10.1109/TII.2014.2349359

Tandon, N., Yadava, G. S., & Ramakrishna, K. M. (2007). A comparison of some condition monitoring techniques for the detection of defect in induction motor ball bearings. *Mechanical Systems and Signal Processing*, *21*(1), 244–256. https://doi.org/10.1016/j.ymssp.2005.08.005

Temporary standards of bandwidth settings for spectral, 2001. *Временные нормативы полосовых уставок для спектрального вибромониторинга газоперекачивающих* *агрегатов ОАО "Газпром", утвержденные 02.04.2001 г. начальником управления по транспортировке газа и газового конденсата ОАО "Газпром" А.З. Шайхутдиновым.*

Terentyev A.N., Sedykh Z.S., V.G. Dubinsky. (1979). *Надежность газоперекачивающих агрегатов с газотурбинным приводом: Reliability of gas compressor units with gas turbine drive.*

Toma, S., Capocchi, L., & Capolino, G.-A [Gerard-Andre] (2013). Wound-Rotor Induction Generator Inter-Turn Short-Circuits Diagnosis Using a New Digital Neural Network. *IEEE Transactions on Industrial Electronics*, *60*(9), 4043–4052. https://doi.org/10.1109/TIE.2012.2229675

Valeev, A., Siraeva, A., & Chen, Y. (2022). Application of clustering algorithms to detect abnormal state of pumping equipment. *Liquid and Gaseous Energy Resources*, *2*(2), 73–79. https://doi.org/10.21595/lger.2022.23079

Valeev A.R. (2020). *Разработка комплексной технологии мониторинга и вибрационной защиты нефтегазоперекачивающего оборудования* [Post-doctoral qualification], Уфимский государственный нефтяной технический университет.

Valeev A.R., O. M. Saubanov (2014). Перспективы внедрения автоматизированной системы контроля фактического состояния насосного оборудования, 19–23.

Voroncov K.V. (2007). *Лекции по методу опорных векторв.*

Wei Zhou, Thomas G. Habetler, & Ronald G. Harley (2007). Stator Current-Based Bearing Fault Detection Techniques: A General Review. In *Diagnostics for Electric Machines, Power Electronics and Drives, 2007. SDEMPED 2007. IEEE International Symposium on: [The 6th IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives : [Hotel Novotel] Cracow (Poland), September 6-8, 2007]* (pp. 7–10). IEEE. https://doi.org/10.1109/DEMPED.2007.4393063

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Yakovlev A.N. (2003). *Введение в Вейвлет-преобразование: Учебное пособие. 104 с.*

Zaritsky S.P. (1987). *Диагностика газоперекачивающих агрегатов газотурбинными приводами: Diagnostics of gas compressor units with gas turbine drives.*

Zhongming Ye, Bin Wu, & A. Sadeghian (2003). Current signature analysis of induction motor mechanical faults by wavelet packet decomposition. *IEEE Transactions on Industrial Electronics*, *50*(6), 1217–1228.

https://www.academia.edu/17545831/Current_signature_analysis_of_induction_moto
r_mechanical_faults_by_wavelet_packet_decomposition

Appendix A

Code for data measuring and preprocessing

This code was designed to measure data from sensors and send to a computer. The code for the Arduino microcontroller is written in Arduino language ("Simplified C++.").

```
unsigned long time;
int N ;
int x[512];
int i;
const int ledPin = 13;
const int pin1 = A2;
const int pin2 = A5;
int incomingByte = 0;
int i_kak_e = 0;
void setup() {
Serial.begin(115200);
pinMode(ledPin, OUTPUT);
digitalWrite(ledPin, HIGH);
}
void loop() {
N = 512;
incomingByte = 0;
// while (incomingByte < 1)
// {
// incomingByte = Serial.read();
// }
time = micros();
for ( i=0; i <= N-1; i++){
if (i % 2 == 0) {x[i] = analogRead(pin1);}
if (i % 2 == 1) {x[i] = analogRead(pin2);}
delay(1) ;
}
time = micros()-time;
for ( i=0; i <= N-1; i++){
Serial.print(x[i]);
Serial.print('*');
}
Serial.print(time);
```

```
Serial.print('*');
Serial.println('*');
delay(100) ;
}
```

Appendix B

Code for data processing and storage based on Python programming language

This code was developed for the purpose of collecting and storing received data from the microcontroller.

```
import serial


def GetOneArrayArd(COMName, COMSpeed):
    ser = serial.Serial(COMName, COMSpeed)
    try:
        s = ser.readline().replace(b'\r\n', b'').decode()
    except TypeError as e:
        print(e)
    ser.close()


    s = s.rsplit('*')
    n = len(s) - 3
    a = []
    b = []
    for i in range(n):
        if i % 2 == 1:
            a.append(int(s[i]))
        else:
            b.append(int(s[i]))
    t = int(s[-3])
    # a - current
    # b - voltage

    return a, b, t
```

Appendix C

Code for machine learning models training based on Python programming language

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from IPython.display import clear_output
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score
import sklearn
from sklearn.metrics import accuracy_score
from sklearn.metrics import make_scorer
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer
from sklearn.model_selection import cross_val_score


# # Чтение всех Excel файлов

# In[ ]:


base_path = r'C:\Users\Lenovo\Desktop\Диплом\EXPER –
копия\Experiment data'


# In[ ]:


freqs = [20,25,30,35,40]
exps = ['rotor', 'rotor_disb', 'wl']
```

```
lists = list(range(100))


# In[ ]:


df0 = None
for freq in freqs:
    for exp in exps:
        path = base_path+'\\'+f'{freq}{exp}.xlsx'

        clear_output(wait=True)
        print(f"Loading file {path}")

        xls = pd.ExcelFile(path)
        for i in lists:
            dfi = pd.read_excel(xls, str(i))

            dfi = dfi[['Current [mA]','Voltage [mV]','Time 
[mcs]']]
            dfi.columns = [f'{freq}{exp}{i}I', 
f'{freq}{exp}{i}V', f'{freq}{exp}{i}t']

            if df0 is None:
                df0 = dfi
            else:
                df0 = pd.concat([df0, dfi], axis=1)

clear_output(wait=True)
print("Loaded!")


# In[ ]:


df0['t'] = df0['20rotor1t'].mean()*df0.index / df0.shape[0] / 
1_000_000


# In[ ]:


total_time = df0['20rotor1t'].mean() / 1_000_000
max_freq = 1 / total_time * df0.shape[0] / 2
min_freq = 1 / total_time
print(total_time)
print(max_freq)
print(min_freq)


# In[ ]:


df0
```

```
#*See how the signal changes over time**

# Voltage change over time

# In[ ]:


df0.plot(x = 't', y = '20rotor3V', grid=True, legend=False,
ylabel = 'measured values', xlabel = 'time, seconds')


# Changes in current strength over time

# In[ ]:


df0.plot(x = 't', y = '20rotor3I', grid=True, legend=False,
ylabel = 'measured values', xlabel = 'time, seconds')



# Conversion of all signals into a spectrum
#
# dff - table with spectra
# In[ ]:


V = np.abs(np.fft.rfft(df0.T)).T / len(df0)


# In[ ]:


V.shape


# In[ ]:


dff = pd.DataFrame(V, columns = df0.columns)


# In[ ]:


dff = dff.drop('t', axis=1)


# In[ ]:


dff['frequency'] = 1/total_time * dff.index #/ dff.shape[0]
```

```
dff

A = np.angle(   np.fft.rfft(   df0.T   )   ).T
A.shape

columns_freq = [str(i)+'_phase' for i in df0.columns]

dffA = pd.DataFrame(A, columns = columns_freq)

dffA

dff2 = pd.concat([dff, dffA],axis=1)

# In[ ]:

dff = dff2


# Voltage spectrum
# In[ ]:

dff.drop(0).plot(x = 'frequency',
                 y = '40rotor0V',
                 grid=True,
                 legend=False,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )


# Current spectrum

# In[ ]:

dff.drop(0).plot(x = 'frequency',
                 y = '40wl0I',
                 grid=True,
                 legend=False,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )

# Comparison of no-load spectra at different frequencies
```

```
# In[ ]:


dff.drop(0).plot(x = 'frequency',
                 y =
['20wl0V','25wl0V','30wl0V','35wl0V','40wl0V'],
                 grid=True,
                 legend=True,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )
plt.legend(['20 Hz','25 Hz','30 Hz','35 Hz','40 Hz'])


** Comparison of voltage spectra in the presence of a load at
different frequencies**


# In[ ]:


dff.drop(0).plot(x = 'frequency',
                 y =
['20rotor0V','25rotor0V','30rotor0V','35rotor0V','40rotor0V'],
                 grid=True,
                 legend=False,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )
plt.legend(['20 Hz','25 Hz','30 Hz','35 Hz','40 Hz'])


# Comparison of spectra at load and unbalance - at different
frequencies

# In[ ]:


dff.drop(0).plot(x = 'frequency',
                 y =
['20rotor_disb0V','25rotor_disb0V','30rotor_disb0V','35rotor_d
isb0V','40rotor_disb0V'],
                 grid=True,
                 legend=False,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )
plt.legend(['20 Hz','25 Hz','30 Hz','35 Hz','40 Hz'])


#** Comparison of current spectra in the presence of a load at
different frequencies**
```

```
# In[ ]:


dff.drop(0).plot(x = 'frequency',
                 y =
['20wl0I','25wl0I','30wl0I','35wl0I','40wl0I'],
                 grid=True,
                 legend=True,
                 ylabel = 'measured values',
                 xlabel = 'frequency, Hz'
                )
plt.legend(['20 Hz','25 Hz','30 Hz','35 Hz','40 Hz'])

# ## Mean value and standart deviation

# In[ ]:


def get_df_mean_std(dff, s, p):

    df_av = None

    for freq in freqs:
        for exp in exps:
            df_ = None

            for i in lists:
                dfi = dff[[f'{freq}{exp}{i}{p}']]

                if df_ is None:
                    df_ = dfi
                else:
                    df_ = pd.concat([df_, dfi], axis=1)

            if s == 'mean':
                df_col = df_.mean(axis=1)
            else:
                df_col = df_.std(axis=1)

            df_col.name= f'{freq}{exp}{p}_{s}'

            if df_av is None:
                df_av = df_col
            else:
                df_av = pd.concat([df_av, df_col], axis=1)

    df_av['frequency'] = 1/total_time * df_av.index
    return df_av


# In[ ]:


df_av = get_df_mean_std(dff, 'mean', 'I')
df_av.drop(0).plot(x = 'frequency',
```

```
                     y = ['20wlI_mean','20rotor_disbI_mean'],
                     grid=True,
                     legend=True,
                     ylabel = 'measured values',
                     xlabel = 'frequency, Hz'
                    )
plt.legend(['Without load','With load'])


# In[ ]:


df_av = get_df_mean_std(dff, 'std', 'I')
df_av.drop(0).plot(x = 'frequency',
                     y = ['20wlI_std','20rotor_disbI_std'],
                     grid=True,
                     legend=True,
                     ylabel = 'measured values',
                     xlabel = 'frequency, Hz'
                    )
plt.legend(['Without load','With load'])


# Let's apply machine learning algorithms to the full set of
raw data without additional processing.
# We will use Linear Regression Classificator and Decision
Tree Classificator models.
# We'll start with all frequencies together, then separately
for frequencies 20, 25, 30, 35, 40 Hz.
#
# We'll also run them separately for voltage, current and both
# In[ ]:


def get_ml(freqs=freqs, exps=exps, lists=lists, vi=['V', 'I'],
aa=[''], k=1):

    df_ml = []

    for freq in freqs:
        for exp in exps:
            for i in lists:
                df_row = None
                index = f'{freq}{exp}{i}'

                for p in vi:
                    #for a in ['','_phase' ]:
                    for a in aa:

                        dfi = dff[[f'{freq}{exp}{i}{p}{a}']]
                        name = dfi.columns[0]

                        dfi = dfi.T

                        dfi.columns = [name[-
```

```
1]+a+'_f'+str(col) for col in dfi.columns]
                        dfi.index = [index]

                        if df_row is None:
                            df_row = dfi
                            df_row['defect'] = 0 if 'wl' in
name else 1
                        else:
                            df_row = pd.concat([df_row, dfi],
axis=1)

                df_ml.append(df_row)

    res = pd.concat(df_ml, axis=0)

    if k==1:
        return  res

    big_rows = []
    nn = []
    for i in range(len(res) // k):
        big_row = res.iloc[i*k:(i+1)*k,:]
        n = big_row.index[0]
        big_row = big_row.max(axis=0)


        nn.append(n)

        big_rows.append(big_row)
    big_rows = pd.concat(big_rows, axis=1)
    big_rows.columns = nn
    big_rows = big_rows.T

    return big_rows


# In[ ]:


df_ml0 = get_ml(k=2)


# In[ ]:


df_ml0


# In[ ]:


for i in df_ml0.columns:
    print(i)
```

```
# In[ ]:


df_ml0['defect']


# In[ ]:


df_ml0.to_excel(r'C:\Users\Lenovo\Desktop\Диплом\dsd.xlsx')


# In[ ]:


def my_f1(y_test, predict):

    defects = 0
    defects_true = 0

    not_defects = 0
    not_defects_true = 0

    for i in range(len(y_test)):

        t = y_test[i]
        p = predict[i]

        if t == 1:
            defects += 1

            if t == p:
                defects_true += 1

        if t == 0:
            not_defects += 1

            if t == p:
                not_defects_true += 1

    sm = 0.00001

    d = defects_true/(defects + sm)
    nd = not_defects_true/(not_defects + sm)

    return d*nd*2/(d+nd)

custom_score = make_scorer(my_f1, greater_is_better=True)


# In[ ]:


models = {'LogisticRegression' :
LogisticRegression(max_iter=1000, random_state=1),
```

```
           'DecisionTreeClassifier' :
DecisionTreeClassifier(random_state=1),
          }

parametrs = {}
parametrs['LogisticRegression'] = {}

parametrs['DecisionTreeClassifier'] = {#'max_depth': range
(3,11, 2),
              'min_samples_leaf': range (2,11,2),
              'min_samples_split': range (2,11,2),
           }


# In[ ]:


model = DecisionTreeClassifier(random_state=1, max_depth = 3)


# In[ ]:


pip install graphviz


# In[ ]:


df_ml = get_ml(freqs=['20'], exps=exps, lists=lists, vi=['V'])

X = df_ml.drop('defect', axis=1) #feature
y = df_ml['defect'] # target
model.fit(X, y)


# In[ ]:


from sklearn import tree


# In[ ]:


from IPython.display import SVG


# In[ ]:


from sklearn import tree
plt.figure(figsize=(40,20))  # customize according to the size
of your tree
_ = tree.plot_tree(model, feature_names = X.columns, rounded =
```

```
True, proportion = False,
                precision = 2, filled = True)
plt.show()
```

List of Figures

List of Tables

Abbreviations

AC (motors)	Alternating current (motors)
AFC	Amplitude-frequency characteristic
AFR	Amplitude-frequency response
ANN	Artificial neural network
CART (method)	Classificator and Decision Tree Classificator
DFT	Discrete Fourier transform
EMF	Electromotive force
ESPRIT	Estimation of signal parameters by rotational invariant techniques
EU	Europe union
FDD	Fault detection and diagnosis
FFT	Fast Fourier transform
FN	False negative
FN	False negative
FP	False positive
FP	False positive
GPU	Gas pumping unit
IDE	Integrated development environment
IMs	Induction machines
IoT	Internet of Things
KDE	Kernel density estimation
M&R	Maintenance and repair
MCSA	Motor current signature analysis
ML	Machine learning

MLP	Multilayer perceptron
MTBF	Mean time between failures
MUSIC	Multiple signal classification
NSs	Neural networks
PC	Personal computer
PCE	Pump-compressor equipment
PdM	Predictive maintenance
RMS	Root-mean-square
SVM	Support vector method
TN	True negative
TP	True positive
TC	Technical condition
VFD	Variable frequency drive
WT	Wavelet transform