AUTOMATED ANOMALY DETECTION IN ENERGY CONSUMPTION

An Undergraduate Research Scholars Thesis

by

KYLE HSU[1], SIXING ZHENG[1]

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Xia Hu

May 2021

Major: Computer Science[1]

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

We, Kyle Hsu and Sixing Zheng, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisors prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

ABSTRACT

Automated Anomaly Detection in Energy Consumption

Kyle Hsu and Sixing Zheng
Department of Computer Science
Texas A&M University

Research Faculty Advisor: Xia Hu
Department of Computer Science
Texas A&M University

There is no world without energy. Dependence on energy continues to dominate everything that we do. It is known that any failure in the production of energy can directly affect thousands of lives. Because of this, data is closely monitored and collected. Our research intends to apply Automated anomaly detection to energy performance data to detect degradations in energy consumption. We use a multivariate time series dataset from a six year period of time at a Combined Cycle Power Plant. Anomaly detection is a data analysis method with the purpose of identifying points in data that do not follow the intended behavior of the dataset. These points can be caused by error, technical faults, or bugs that have potentially devastating impacts if unnoticed. Anomaly detection has become more flexible as more methods of data processing, feature analysis, and detection have become available. Although the techniques of anomaly detection have drastically improved in recent years, there has been little research done on automated anomaly detection. Building an anomaly detection system requires an expert to manually select features such as data pre-processing methods and feature analysis methods in order to construct an anomaly detection pipeline that is suitable for the dataset. This method is

very costly and can be done with a machine learning approach. With the incorporation of machine learning, Automated Anomaly Detection has the ability to build an optimal pipeline according to the types of the dataset. Instead of wasting time and money manually building possibly unreliable anomaly detection algorithms, the process is simplified by just feeding in the desired dataset to detect anomalies. The system would process the dataset and get information by running the system on the dataset. According to the information, the system would pick different components and build pipelines to check for the accuracy until the best optimized pipeline is generated for the dataset. Our Anomaly detection system construction is built by Time Series Outlier Detection System (TODS) which implements modern machine learning principles to construct an optimal anomaly detection system for our time series dataset. TODS focuses on data processing, time series processing, feature analysis, and detection algorithms to construct an outlier detection system. Utilizing automated anomaly detection methods in monitoring energy consumption can help energy consumers quickly find and fix the degraded parts to minimize consumption and provide more safety to users. The requirement of an expert in energy performance is no longer needed. The cost of constructing anomaly detection pipelines for industries would significantly be lower as well.

ACKNOWLEDGEMENTS

Contributors

I would like to thank my faculty advisor, Dr. Hu, and our research leads, Kwei-Herng Lai, and Daochen Zha for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

The resources used for automated anomaly detection in energy consumption were provided by the Data Lab at Texas A&M. The analyses depicted in automated anomaly detection in energy consumption were conducted in part by Kyle Hsu and Sixing Zheng. The energy dataset was provided by the UCI Machine Learning Repository.

All other work conducted for the thesis was completed by the students independently.

Funding Sources

We received no funding for this research.

NOMENCLATURE

B/CS	Bryan/College Station
TODS	Time Series Outlier Detection System
CCPP	Combined Cycle Power Plant

1. INTRODUCTION

1.1 Anomaly Detection

Anomaly detection is the detection of faults, uniqueness, or oddities in patterns or trends of data. These points are also known as outliers since they do not belong in the dataset and drastically influences the data statistics. Anomaly detection is split into three different types: contextual anomalies, collective anomalies, and point wise anomalies. Contextual anomalies are anomalies that occur when there is an unusual change in an expected pattern. Collective anomalies occur when objects are grouped together, and one or more objects do not belong to a specific group. Point wise anomalies are points in a dataset that are drastically different from the other data points. In other words, these specific data points stand out as unique because their values do not make sense when compared to all other data points. Our research focuses on applying a machine learning approach to point wise anomalies.

1.2 Applying a Machine Learning Approach to anomaly detection

Over the recent years, anomaly detection has grown to have a huge impact in data analysis. However, it is rare to see automated anomaly detection with the use of modern-day principles of Machine Learning. This is largely due to the difficulty in building an optimal system for detection with different types of data. Time Series Outlier Detection System (TODS) aims to overcome that barrier and provide an optimal system that suits the input data.

1.2.1 Time Series Outlier Detection System (TODS)

Time Series Outlier Detection System (TODS) is a full stack machine learning system developed by the DATA Lab at Texas A&M that detects anomalies in multivariate time series data. TODS features a modern machine learning approach to the selection of many data

preprocessing methods, feature analysis methods, and outlier detection methods that are utilized to construct an outlier detection system for any given data. TODS excels at point-wise detection, system detection, and pattern-wise detection. TODS also provides modules including data processing, time series processing, feature analysis, detection algorithms, and reinforcement module to make machine learning based anomaly detection algorithms available. These functionalities from the mentioned module include data preprocessing for general purposes, time series data smoothing and transformation, extraction features from time and frequency domains, various detection algorithms, and involving human expertise to calibrate the system. Our code is written in Python and incorporates modern day machine learning libraries such as Scikit, D3M, etc.

1.3 Applying TODS to Energy Consumption

TODS is still expanding and growing more accustom to processing data in different fields. The target is to expand TODS into the energy industry. The time series energy dataset that is used is from the UCI Machine Learning repository. This dataset contains multivariate data over a six year period of time at a Combined Cycle Power Plant (CCPP) and has over nine thousand data points. The data is split into Temperature, Ambient Pressure, Relative Humidity, and Exhaust Vacuum. Each of these variables contributes to the overall output of the power plant. A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another [4]. This dataset is a rough estimate of what to expect from most combined cycle power plants.

1.3.1 Methodology and Results (nee help with formatting)

We start by calibrating TODS to accept the dataset. The important calibrations are the target indexes and the metric used. Once calibrated, we feed in the dataset. TODS will determine the best methods of processing, feature analysis, and detection algorithms to apply to this dataset. TODS will then report any outliers and the percentage accuracy that is predicted for the given dataset. We will manually construct an optimal outlier detection system and compare the results between the system constructed by TODS and the optimal system that was manually constructed. This process is repeated to test accuracy between runs and to draw accurate conclusions.

2. METHODS

TODS uses automated machine learning to assist in anomaly detection. The effectiveness of applying TODS to energy consumption data is measured by the accuracy of the outlier detection system that is constructed by TODS. We decided to target data from an energy power plant. By applying a dataset from a powerplant, we will be able to determine if energy performance degradation can be detected with the use of an automated machine learning approach to anomaly detection. The results will provide helpful insight in the application of TODS to energy consumption.

2.1 Choosing the Right Dataset

We begin by targeting datasets that are specifically from power plants. Since the source of many types of energy originate from these plants, we determined that anomaly detection has its best impact when detecting energy degradation from the source. TODS is constructed for the application of anomaly detection towards time series data. That means our dataset must be a time series dataset. The energy performance dataset that we settled on was a multivariate time series data from a CCPP which contains over nine thousand data points collected over a six year time period. More specifically, this dataset contains variables from Gas Turbines and Steam Turbines including temperature (Figure 1.1), ambient pressure (Figure 1.3), relative humidity (Figure 1.4), exhaust vacuum (Figure 1.2), and the electrical energy output (Figure 1.5). We expect most combined cycle powerplants to behave the same way.

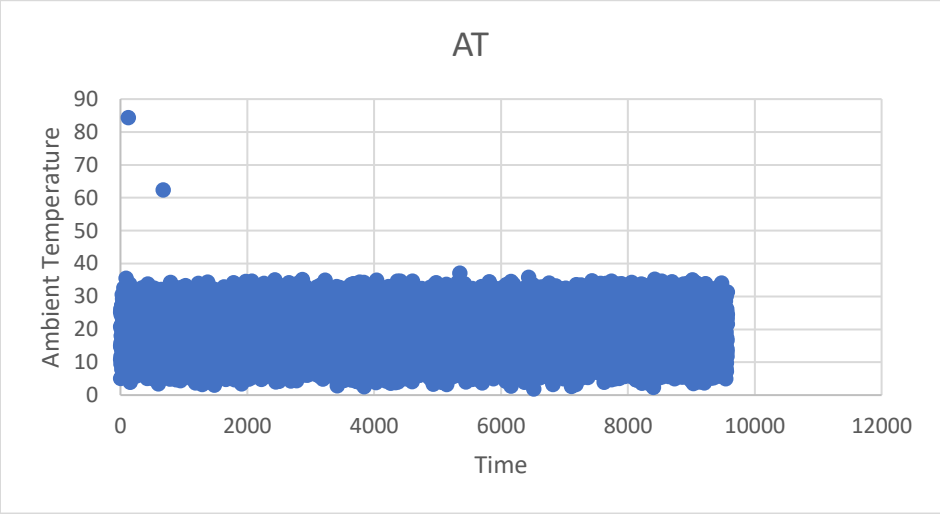

Figure 1.1: Ambient Temperature over a six year time period

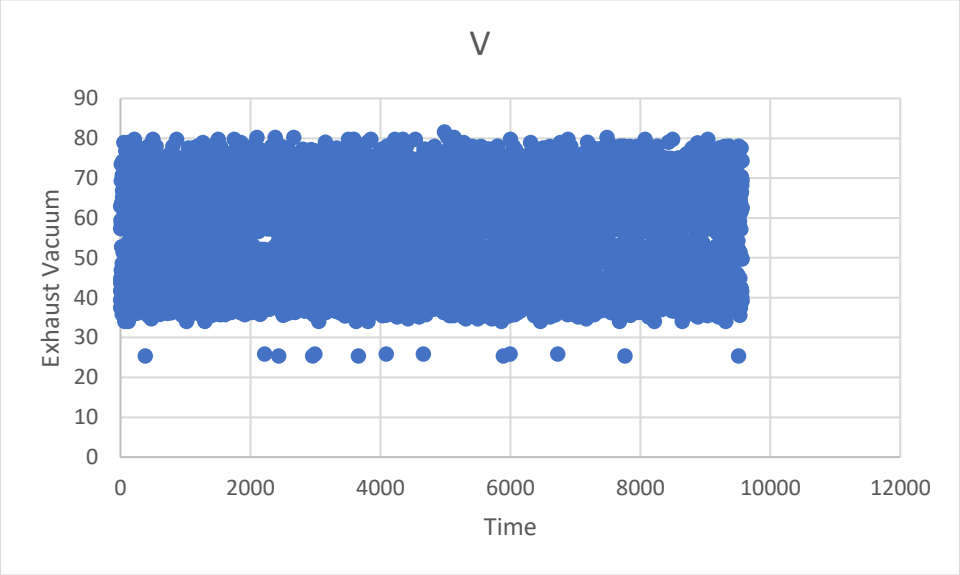

Figure 1.2: Exhaust Vacuum over a six year time period

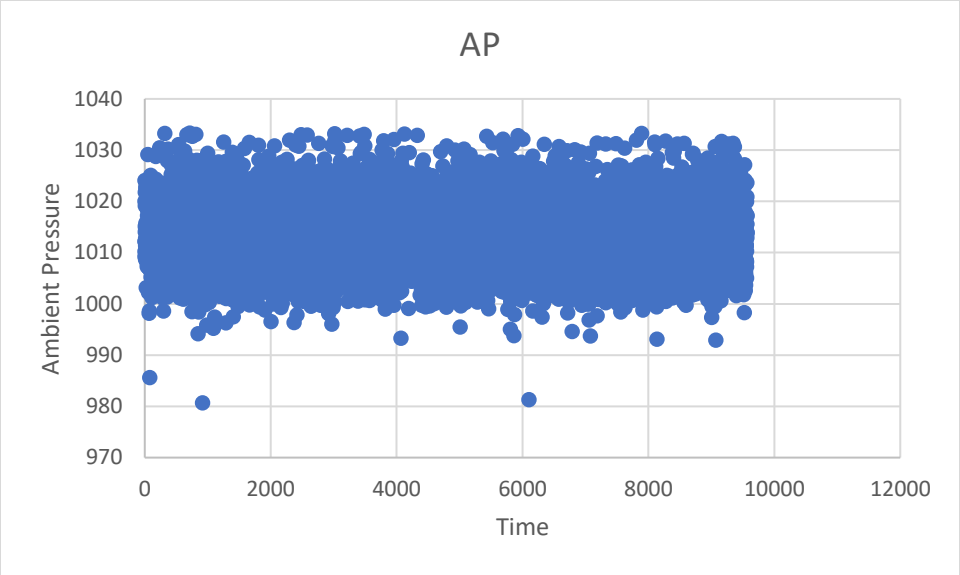

Figure 1.3: Ambient Pressure over a six year period of time

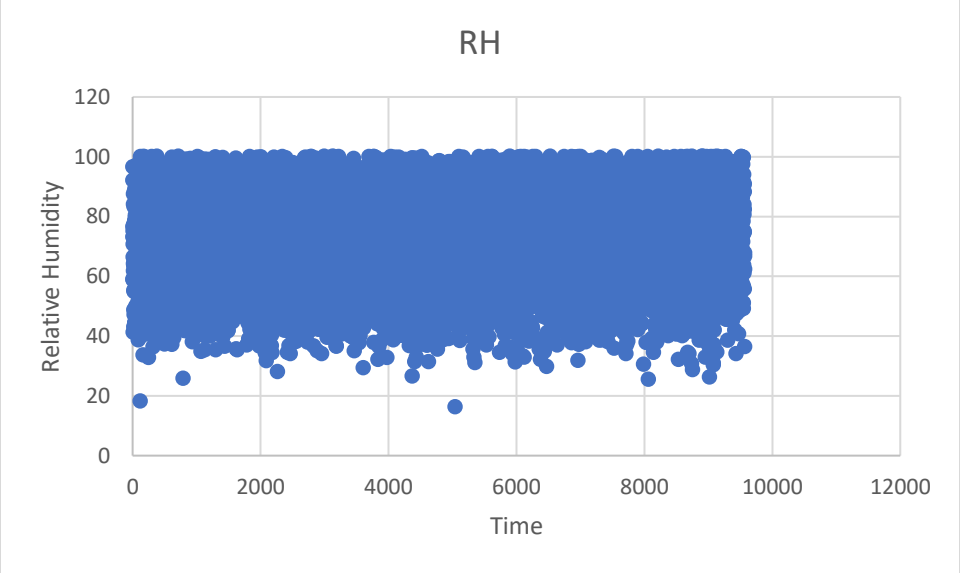

Figure 1.4: Relative Humidity over a six year period of time

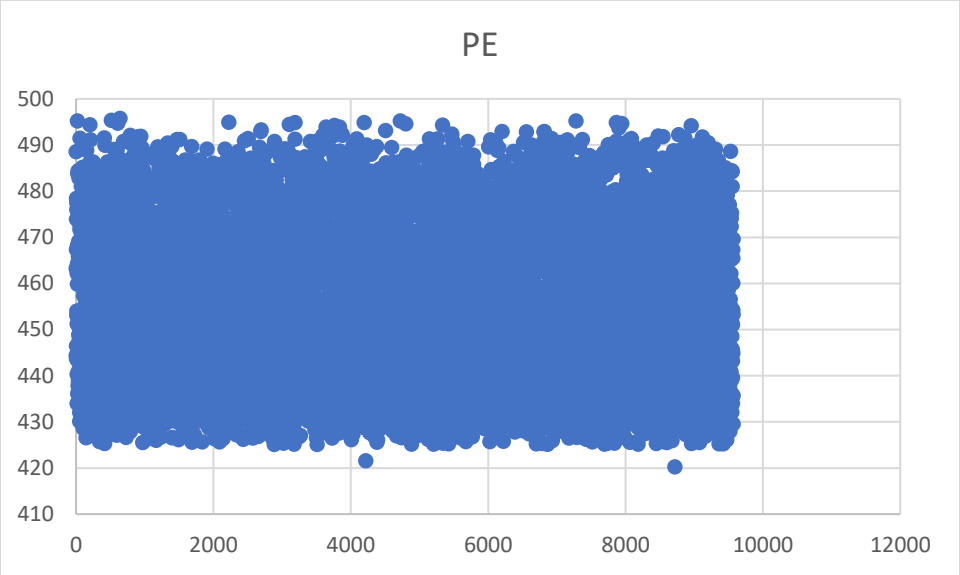

Figure 1.5: Electrical Output over a six year period of time

Each of the tables (Figures 1.1-1.5) visually displays the data points we will analyze and feed into TODS. It is easy to tell from visuals alone which points are potential outliers. However, the constructed anomaly detection pipelines will guarantee which points are actual anomalies.

2.2 Calibrating TODS

After a dataset has been selected, we manually change static variables to prepare TODS for the dataset. The first variable to be changed is the target index. This variable must be changed depending on the amount of data and the type of data. Specifically, in our selected dataset, we manually change TODS to target the following types of data in our dataset: Temperature, Ambient Pressure, Relative Humidity, and Exhaust Vacuum. Next, we select the metric at which TODS will evaluate the dataset. For our dataset, we selected the metric: F1_MACRO. Lastly, we allow TODS to provide its automated machine learning approach to construct our anomaly detection pipeline.

2.3 Feeding in the Dataset

While TODS does support system-wise detection and patter-wise detection, the detection method used will be point-wise detection. This method of detection will return individual time points that are regarded as outliers. Furthermore, the accuracy of detection will be produced as a percentage. TODS uses modules to build an optimal machine learning outlier detection system for the dataset. TODS will either smooth or transform the data using its data smoothing modules. Next, TODS will select an optimal feature analysis (extraction) method and detection algorithm. Once all the methods have been selected, TODS constructs an anomaly detection pipeline and applies the detection system to our selected dataset

2.4 Manually Constructing an Optimal System

Before a machine learning approach was implemented, anomaly detection systems were created manually. The best way to test the accuracy of the system that TODS constructs is to create an optimal system manually. Since we are performing a point wise detection, we decided to approach anomaly detection with a standard deviation system and a box plot system. The standard deviation system calculates mean and standard deviation of each variate and detects outliers by comparing the standard deviation to the mean. On the other hand, boxplots are another system which the median is determined. From there, we then find the quartiles and calculate the interquartile range. Lastly, any point found to be outside the given bounds are considered an outlier.

2.5 Drawing Conclusions

Both the manual system and the automated system will return the exact data points that are regarded as outliers. It is important to compare the methods of data processing and feature extraction because different methods will yield different results. When running multiple

instances of detection, we can manually change the methods of data processing and feature extraction to obtain the most accurate anomaly detection system. We then compare the optimal system to the system that was constructed by TODS. Any significant difference between the optimal system and the automated system will give insight on the effectiveness of TODS.

3. RESULTS

Once the tests have been run, we can compare the accuracies of the system developed by TODS in comparison to a manually designed optimal system. The two factors that we can draw conclusions on are the modules used for each system and the accuracy of outlier detection inside the system. This will help us understand the effectiveness of TODS in detecting performance issues based on multivariate time series energy data.

3.1 System Breakdown

Since the chosen time series dataset is multivariate and contains data points, we must measure each variate individually. TODS will be fed in the data from each of the variates in the form of columns. The data points are written inside an excel sheet with each column representing a different variate. The metric of measurement that we calibrated TODS to be a macro average F1 score from the Scikit learn module. This means that for each variate, we compute an F1 score and returns an unweighted average of the score. The average is then used to compare to the data points to determine which points in the dataset will be considered an outlier. On the other hand, The manually constructed dataset will utilize standard deviations inside of it's system. Each point is added to a sum which then computes the average and the standard deviation. From there we iterate through each data point again to find the number of standard deviations each point is from the calculated mean. We find that if the standard deviation of each point is greater than three, it is highly considered to be an outlier. Three standard deviations away from the mean results in an outlier. Furthermore, we also constructed a manual system using boxplots to further test the accuracy of a manually constructed system. Our boxplot focuses on the interquartile range in order to determine outliers. Once the median is determined, a box plot can be

constructed by determining which values are exactly between twenty five percent and seventy five percent of the dataset. In other words, the two values, quartile one and quartile three contain ranges of twenty five percent and seventy five percent of the dataset respectively. The interquartile range is found by subtracting the third quartile value by the first quartile value. We can consider a point as an outlier if it is greater than the third quartile plus 1.5x the interquartile range or if it is less than the first quartile minus 1.5x the interquartile range.

3.2 Results of multiple trials

The two systems: automatic and manual yielded results that are very similar. In fact, it was found that TODS produced an accuracy of 100 percent in anomaly detection. This indicates that every outlier that was found was a true outlier. Two anomalies were detected in ambient pressure, thirteen anomalies were detected in exhaust vacuum data, three anomalies were detected in ambient pressure, seven anomalies were detected in Relative Humidity, and two anomalies were detected in energy output. Similarly, the manual system produced results that had also had an accuracy of 100 percent. The manual system detected the same data points that TODS also detected. However, we noticed that TODS also had an 87.5 percent chance of false positives. This means that 87.5 percent of the time, a data point fed into TODS will yield a false result. The dataset was fed into TODS and executed multiple times in order to ensure that a single execution was not by chance. Furthermore, the same number of execution times was also applied to the manual system.

4. CONCLUSION

4.1 Impact of Anomaly Detection

Failure inside gas turbines and steam turbines can cause an entire power plant to shut down. Failure in power plants results in failures to produce energy. Energy is responsible for the continuation of daily life and is found in every aspect of modern society. Each of these machines have different variables that could cause failures. Variables such as Ambient Temperature, Ambient pressure, etc. could cause issues in the turbines and engines. It is important that any potential failure in these machines can be detected as early as possible in order to maximize energy production. Our machine learning approach to anomaly detection allows power plants to feed in their time series data into the system which constructs an optimal pipeline that accurately detects anomalies within the dataset. Our approach significantly reduces the cost of detection. In the past, energy degradation is usually found after there is a clear issue with the engines. Slowly as technology progressed and patterns in data were becoming more accurate, engine failures did not occur as often. In modern times, enough data has been collected by experts to be able to predict when an engine failure is possible. However, data analysis in the form of anomaly detection has always been done manually by an expert in the field. The results produced by the manually constructed system was overwhelmingly positive. The one hundred percent accuracy indicates that the chance that a detected anomaly is not an anomaly is zero percent. Although the results were highly accurate, we strongly believe that a manual system can be replaced by an automated system like our system with TODS and achieve exceptional results.

4.2 Results of an Automated Approach

Like the manually constructed system, our automated approach to anomaly detection also yielded an accuracy of one hundred percent. The system is one hundred percent confident that every anomaly detected is an anomaly. When comparing the manual system to the automated system, was found that the anomalies found in the dataset were the same. However, the existence of false positives introduces error into the systems. Our result of an 87.5 percent chance of a false positive is extremely high. Although our system is highly accurate, there is a chance that the points our system detected may not be an anomaly at all. This error may not be due to our system itself since the manual system also produced similar results. In fact, it is more likely that the false positives are due to the measurements within the dataset itself. We strongly believe that false positives are insignificant as long as the accuracy of the system itself is high. The one drawback is that our system was only tested on one dataset. This means that other datasets will yield different results. Results can further vary depending on the modules used and the accuracy. We have yielded very positive results and definitely recommend applying TODS to more powerplant datasets to further explore the accuracies of a machine learning approach to anomaly detection in energy consumption.

REFERENCES

[1] Araya, D. B., Grolinger, K., ElYamany, H. F., Capretz, M. A., & Bitsuamlak, G. (2017). *An ensemble learning framework for anomaly detection in building energy consumption* (pp. 191-206). ScienceDirect. doi:https://doi.org/10.1016/j.enbuild.2017.02.058.

[2] Djurdjanovic, D.; Lee, J.; and Ni, J. 2003. *Watchdog agentan infotronics-based prognostics approach for product performance degradation assessment and prediction.* Adv. Eng. Inform. 17(3-4):109–125.

[3] Chou, J., & Telaga, A. S. (2014). *Real-time detection of anomalous power consumption* (Master's thesis, 2014). ScienceDirect. doi:https://doi.org/10.1016/j.rser.2014.01.088

[4] Heysem Kaya, Pınar Tüfekci , Sadık Fikret Gürgen (2012). *Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering* ICETCEE 2012, pp. 13-18 (Mar. 2012, Dubai)