VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI, KARNATAKA - 590 018.

PROJECT REPORT ON

Pump Anomaly Detection Using Machine Learning Techniques

Submitted in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

MANTHAN PRASAD [1JS20CS088]

NISHANTH A R [1JS20CS102]

RAHUL R [1JS20CS120]

RAHUL R [1JS20CS121]

Under the guidance of

Ms. Shanthala K V,
Assistant Professor,
Dept of CSE, JSSATE,
Bengaluru

Department of Computer Science and Engineering

JSS ACADEMY OF TECHNICAL EDUCATION, BENGALURU

2023 – 2024

JSS MAHAVIDYAPEETHA, MYSURU

JSS Academy of Technical Education

JSS Campus, Uttarahalli Kengeri Main Road, Bengaluru – 560060

Department of Computer Science and Engineering

CERTIFICATE

Certified that the project work entitled "Pump anomaly detection using Machine learning Techniques" carried out by Manthan Prasad (1JS20CS088), Nishanth A R (1JS20CS102), Rahul R (1JS20CS120), Rahul R (1JS20CS121) in partial fulfilment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visveswaraiah Technological University, Belgaum during the year 2023-2024. It is certified that all corrections and suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

----------	----------	----------
Signature of the Guide	**Signature of the H.O.D**	**Signature of the Principal**
Ms Shanthala K V	**Dr. Mallikarjuna P B**	**Dr. Bhimasen Soragaon**
Assistant Professor,	**Professor & Head**	**Principal,**
CSE, JSSATE, Bengaluru	**CSE, JSSATE, Bengaluru**	**JSSATE, Bengaluru**

Name of the examiners Signature with date

1.

2.

ACKNOWLEDGEMENT

We express our humble gratitude to His **Holiness Jagadguru Sri Sri Sri Shivarathri Deshikendra Mahaswamiji** for showering us with blessings and guiding us towards a successful career.

The completion of any project is a collaborative effort involving many individuals. We have been fortunate to receive substantial help and support from various quarters throughout this project. With heartfelt gratitude, we acknowledge all those whose guidance and encouragement have contributed to our success.

We are deeply thankful for the invaluable guidance, timely assistance, and gracious support of our guide, **Ms. Shanthala K V**, Assistant Professor in the Department of Computer Science, who has been instrumental in every aspect of our work. Our gratitude also extends to **Dr. P B Mallikarjuna**, Professor and Head of the Department of Computer Science and Engineering, for the facilities and support provided. We sincerely thank our beloved principal, **Dr. Bhimasen Soragaon**, for his unwavering support in our academic endeavors.

Additionally, we would like to acknowledge **Pump Academy** for providing the essential datasets for our project. We recognize that this project would not have been possible without their invaluable assistance and the guidance of **Dr. Ravindra K Shetty**.

Lastly, we express our heartfelt thanks to all the teaching and non-teaching staff of the CSE department and our friends for their help, motivation, and support.

Manthan Prasad (1JS20CS088)
Nishanth A R (1JS20CS102)
Rahul R (1JS20CS120)
Rahul R (1JS20CS121)

ABSTRACT

In contemporary production processes, human operators traditionally oversee operations, yet may not always intervene promptly in case of breakdowns. The emergence of machine learning and the Internet of Things (IoT) has facilitated the development of automated solutions. These technologies enable the monitoring of production processes and the implementation of predictive maintenance systems, potentially reducing maintenance frequency and failures. Motors operating at optimal speeds with natural frequencies can indicate anomalies leading to pump underperformance. Predictive maintenance, a strategy utilizing machine learning, predicts equipment failure probabilities. Industrial equipment, like water pumps, generates diverse signals during operation. By installing sensors for data collection, machine learning can discern normal and abnormal behavioral patterns. This proactive approach aims to minimize downtime, enhance efficiency, and prolong equipment lifespan. This project centers on predictive maintenance for pump systems, employing machine learning and forecasting models. The dataset, sourced from Pump Academy, undergoes meticulous preprocessing to manage missing values, outliers, and feature engineering. Exploratory data analysis (EDA) illuminates' relationships between voltage, power, and current data. Various machine learning models including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and XGBoost classify pump conditions into normal, alarm, and shutdown states. Evaluation metrics like accuracy, precision, and recall guide model selection. Moreover, forecasting models such as Prophet and ARIMA predict future pump conditions. Rigorous parameter tuning and validation ensure accuracy. The outcomes demonstrate the efficacy of classification and forecasting models in predicting and identifying pump system conditions. This approach holds significant potential for predictive maintenance in industrial settings, bolstering reliability and minimizing downtime.

Table of Contents

Chapter Title **Page No.**

List of Figures

Figure No.	Name of the Figure	Page No.

List of Tables

Table No.	Name of the Table	Page No.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

Predictive maintenance, generally known as "monitoring system" or "risk-based maintenance," has been studied in a number of recent journals. It defines the intelligent monitoring equipment to avoid future failures or consequences. Predictive Preservation has evolved from the basic strategy of visual data evaluation to automated systems based on advanced methodologies based on Pattern Recognition, Machine Learning (ML), and other technologies. Many industry sectors now have a viable alternative for detecting and gathering sensitive information from a range of equipment, particularly machines, where human sight or hearing may be limited. Predictive preservation and integrated sensors can help you avoid unnecessary equipment part replacements, decrease machine downtime or failure, identify the root cause of a problem, and save money. In contrast to conservative preventative preservation, predictive preservation schedule activities are based on collecting datasets from various sensors and performing analysis algorithms.

1.1 Motivation

Detecting anomalies in industrial pump systems is crucial for minimizing downtime, reducing maintenance costs, ensuring safety, and optimizing operational efficiency. Unplanned downtime due to pump failures can lead to significant financial losses, making proactive anomaly detection essential for timely maintenance interventions. Additionally, preserving equipment integrity through early anomaly detection extends its operational lifespan and enhances safety for personnel and assets. Furthermore, anomaly detection enables process optimization by providing insights into performance parameters, ensuring regulatory compliance, and serving as a cornerstone for predictive maintenance strategies, ultimately driving operational resilience and efficiency in industrial environments.

1.2 Existing System

Traditionally, industries have relied on rule-based systems to detect anomalies in pump operations. These systems set predefined thresholds and monitor conditions to spot any deviations from the norm. However, their rigid nature often fails to capture the subtle variations in operational patterns accurately. This leads to frequent false alarms and delays in identifying actual anomalies, highlighting the shortcomings of these conventional methods. With industrial processes becoming more complex, there's a growing need for

smarter, data-driven approaches. It's clear that we must move towards adaptive and intelligent methodologies to address the evolving challenges in pump maintenance and operations.

1.3 Limitations of the Existing System

Traditional methods of anomaly detection have limitations in detecting complex anomalies and adapting to evolving data patterns. These methods often rely on manual rule-based approaches or statistical techniques, which may not be effective in detecting anomalies in large and complex datasets. Additionally, traditional methods may not be able to adapt to changing data patterns, making them less effective in predicting future anomalies.

1.4 Proposed System

This project focuses on predictive maintenance for pump systems using a combination of machine learning and forecasting models. Machine learning models including K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forest and XGBoost are employed to classify pump conditions into normal, alarm, and shutdown states. Furthermore, forecasting models such as Prophet and ARIMA are implemented to predict future pump conditions. Model parameters are tuned and validated to ensure accuracy.

1.5 Advantages of the Proposed System

The proposed system combines machine learning and forecasting models to offer several advantages. It enhances accuracy through the utilization of KNN, SVM, Random Forest, and XGBoost, efficiently classifying pump conditions into normal, alarm, and shutdown states. With inherent flexibility, it adapts to varying operational patterns without manual intervention, ensuring robust performance. Additionally, the system's integration of machine learning and forecasting enables effective anomaly detection by analysing historical data and predicting future conditions accurately. This proactive approach enables timely maintenance interventions, reducing downtime and optimizing equipment uptime, ultimately enhancing the reliability and efficiency of pump systems.

CHAPTER 2
LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

Literature surveys are essential in understanding the current state of research and identifying gaps for further investigation. In the context of pump anomaly detection, integrating computer science and mechanical engineering knowledge is vital. This fusion enables the development of sophisticated machine learning models that can effectively predict and diagnose pump failures. The literature survey provides insights into existing methodologies, highlighting their strengths and weaknesses, and guides researchers in refining their approaches to enhance predictive maintenance systems. This comprehensive understanding is crucial for advancing technology and ensuring the reliability and efficiency of pump operations.

2.1 Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs)

Publication year: May 2022

Author: Abdalla, Ramez, Samara, Hanin, Perozo, Nelson, Paz, Carlos, Jaeger, Philip.

Journal Name: ACS Omega 7(6)

Summary: This paper introduces a novel machine learning approach designed for predictive maintenance of Electrical Submersible Pumps (ESPs) within the petroleum industry. Through the fusion of principal component analysis and extreme gradient boosting trees, the model successfully predicts ESP failures up to seven days in advance, achieving an F1-score exceeding 0.71. The methodology encompasses various stages, including data collection, preprocessing, feature extraction, and model training, all aimed at proactively identifying failure events to enhance operational efficiency and minimize downtime and maintenance costs. While the model demonstrates high precision, it also reveals lower recall, indicating that not all pre-failure days are identified as alarms. Model validation results in a mean AUC of 0.95, with precision and recall values of 0.8 and 0.6, respectively, for preworkover and workover classes. Additionally, the paper delves into the utilization of XGBoost and machine learning algorithms for fault detection and classification in oil wells, advocating for the adoption of deep learning across various industrial processes. Furthermore, the paper proposes a predictive model leveraging sensor measurements and

XGBoost to forecast pumping conditions, achieving an impressive ROC AUC of 0.95. Notably, the paper underscores the efficacy of principal component analysis for anomaly detection in ESPs, emphasizing the critical role of data preprocessing, feature engineering, and model training. It concludes by suggesting future research directions, particularly in addressing imbalanced datasets through techniques such as oversampling with SMOTE.

2.2 An Industrial Case Study Using Vibration Data and Machine Learning to Predict Asset Health

Publication year: 2018

Author: Amihai, Ido, Gitzel, Ralf, Kotriwala, Arzam, Pareschi, Diego, Subbiah, Subanatarajan, Sosale, Guru.

Journal Name: IEEE

Summary: This case study explores the application of machine learning techniques, specifically Random Forest and Persistence algorithms, for predicting asset health in industrial plants using real-world data. The findings indicate that significant predictions regarding Key Condition Indicators (KCIs) can be made up to 7 days in advance, facilitating the early detection of potential failures in plant assets. However, the research underscores the challenges inherent in data preparation, highlighting the critical importance of addressing data quality issues. Validation through root cause analysis confirms the accuracy of the predictions in identifying asset failures. Despite the success in predicting KCIs, the study acknowledges limitations in detecting unexpected events through automated predictions, suggesting the necessity for further research to enhance prediction quality and explore alternative machine learning algorithms. By gathering vibration data from 30 pumps over a 2.5-year period, the study demonstrates the practical application of machine learning in an industrial setting and underscores the significance of reliable predictions in preventing asset failures. Overall, the case study underscores the potential of machine learning in predictive maintenance while acknowledging the complexities and challenges associated with working with real-world industrial data.

2.3 Predictive Maintenance Using Machine Learning on Water Pump

Publication year: 2022

Author: Dr. Sharda Chhabria, Rahul Ghata, Varun Mehta, Ayushi Ghosekar, Manasi Araspure, Nandita Pakhid.

Journal Name: IRJMETS

Summary: This research article presents the development and implementation of a predictive maintenance system for water pumps utilizing machine learning techniques and sensor nodes. The system architecture encompasses data acquisition, analysis, state detection, health evaluation, and prognosis, integrating sensor nodes, a Wi-Fi MCU module, cloud services, and graphical user interfaces. By leveraging historical machine data to train a machine learning model, the system aims to identify faults in water pump systems early, thereby preventing breakdowns and optimizing maintenance efficiency. Various sensors collect data on machine health, which is transmitted to a cloud platform for analysis. Utilizing machine learning algorithms such as Random Forest, the system can predict potential failures and prompt maintenance alarms. The paper underscores the significant cost-saving potential and performance enhancement benefits associated with predictive maintenance. Additionally, it offers references for further exploration of the topic. Overall, the research illustrates the efficacy of predictive maintenance systems in enhancing operational efficiency and mitigating costly breakdowns in industrial machinery, particularly water pumps.

2.4 Performance Analysis of Centrifugal Pumps Subjected to Voltage Variation and Unbalance

Publication year: 2008

Author: P. G. Kini, R. C. Bansal and R. S. Aithal

Journal Name: IEEE

Summary: The study investigates the influence of voltage variation and imbalance on centrifugal pump systems powered by induction motors, highlighting the importance of considering these factors in motor-pump system selection to optimize performance.

Through a test example, the authors demonstrate the adverse effects of voltage imbalance on system efficiency, focusing on motor torque, power output, and efficiency under different voltage conditions. The analysis underscores the necessity of accurate estimation to enhance energy efficiency and emphasizes the significance of thoroughly studying process requirements before system selection to maximize available conditions. Acknowledgment is extended to Dr. P. N. Sreedhar for his support in the study. The research is supported by various references and research findings, emphasizing the imperative of precise estimation and analysis to enhance pump system energy efficiency. Findings indicate that deviations from rated supply specifications significantly affect pump efficiency, resulting in electrical energy wastage. Furthermore, system efficiency experiences notable reduction under voltage variation and imbalance conditions, highlighting the pivotal role of maintaining stable voltage levels for optimal pump system performance.

2.5 Fault Detection for Circulating Water Pump Using Time Series Forecasting and Outlier Detection

Publication year: 2017

Author: M. Sanayha and P. Vateekul

Journal Name: IEEE

Summary: This paper introduces a two-stage model for fault detection in circulating water pumps, integrating time series forecasting and outlier detection techniques. In the first stage, the model employs ARIMA to forecast trends in sensor data, while the second stage focuses on classifying failure modes based on the predicted sensor values. Through experiments conducted using data from eight sensors collected over a year, the study illustrates that the proposed algorithm surpasses existing methods for fault detection in circulating water pumps. The research underscores the significance of data mining methods for predictive maintenance in pump systems, particularly emphasizing the utilization of time-series prediction models such as REG-NN and ARIMA for anticipating equipment malfunctions. Comparison between the two models reveals that ARIMA demonstrates superior performance in terms of accuracy and execution time. Furthermore, multivariate outlier detection is utilized to identify faults in the predicted data, enhancing the model's capacity to detect equipment breakdowns or malfunctions in advance. Overall, the proposed

method shows promising results in assisting maintenance engineers by facilitating early detection of potential issues, enabling them to prepare necessary resources for inspection or maintenance tasks. The study leverages the R language for data analysis and visualization purposes.

2.6 Vibration analysis for bearing fault detection and classification using an intelligent filter

Publication year: 2014

Author: Jafar Zarei, Mohammad Amin Tajeddini, Hamid Reza Karimi

Journal Name: ELSEVIER

Summary: The paper introduces three distinct approaches for fault detection and classification in induction motors, all relying on artificial neural networks (ANNs). The first method utilizes a prototype network to identify patterns and detect faults in the motors, exhibiting superior performance compared to ANFIS, particularly in scenarios with poor signal quality. This approach proves to be versatile and applicable to various fault types. The second approach involves employing a neural network filter to isolate non-bearing fault components in vibration signals, followed by another neural network for fault classification. This method primarily focuses on time-domain features analysis and does not necessitate intricate calculations or detailed knowledge of machine parameters. Experimental findings validate the effectiveness of this approach in detecting bearing defects in induction motors. Lastly, a novel algorithm is introduced for fault detection and classification, specifically targeting bearings. By utilizing a neural network filter to enhance fault detection accuracy by filtering out non-bearing fault components from vibration signals, followed by time-domain feature extraction for fault classification, this algorithm demonstrates promising results even in noisy industrial environments. Overall, these intelligent methods offer effective solutions for fault detection and classification in induction motors and bearings, showcasing their potential for enhancing operational reliability and efficiency in industrial settings.

CHAPTER 3
SYSTEM ANALYSIS

CHAPTER 3

SYSTEM ANALYSIS

System analysis involves understanding and specifying in detail what an information system should do. It entails identifying the system's requirements, analyzing current operations, and designing solutions that meet business needs. In the context of pump anomaly detection using machine learning, system analysis includes gathering detailed requirements about the data to be collected, the features to be extracted, and the algorithms to be employed. This phase is critical to ensure that the developed system aligns with the operational objectives and effectively addresses the identified challenges in pump maintenance.

3.1 PUMP SYSTEMS AND SENSOR TECHNOLOGIES

Pumping systems are vital in various industrial processes, with different types of pumps designed to meet specific production needs. Among these, centrifugal pumps are the most prevalent in industrial applications. These pumps fall under the category of dynamic pumps and can be further classified into axial flow pumps and radial flow pumps. They come in various configurations such as single or multi-stage, horizontal or vertical alignment, and with impellers that may be open, semi-open, or closed. A centrifugal pump is a rotodynamic hydraulic machine that converts mechanical energy from the impeller into kinetic or pressure energy, which is then imparted to an incompressible fluid. The fluid is drawn in through a suction pipe into the center of the impeller. The impeller, equipped with a series of blades, moves the fluid by centrifugal force toward the discharge pipe. In the case of multi-stage pumps, the fluid may pass through multiple impellers before being discharged.

Centrifugal pumps are essential in maintaining the flow of fluids in industrial processes. Their design and operation make them suitable for various applications, from simple water supply systems to complex chemical processing setups. The versatility in their design allows them to handle different fluid properties and flow requirements.

In this project, we focus on the critical role played by a centrifugal water pump used by the Bangalore Water Supply and Sewerage Board (BWSSB). This pump operates continuously to ensure an uninterrupted water supply, highlighting its importance in maintaining essential public services. The continuous operation of the BWSSB's centrifugal pumps

underscores the reliability and efficiency required of such systems in urban water management. These pumps must be robust enough to handle the demands of a large-scale water supply network, providing a consistent flow of water to meet the needs of the city's population.

3.2 Overview of Pump Systems

Within the scope of this project, our attention is directed towards the centrifugal pumps operational at the TK Halli pumping station under the auspices of the Bangalore Water Supply and Sewerage Board (BWSSB). These centrifugal pumps, featuring power ratings spanning from 1200 kW to 1500 kW, constitute indispensable assets within the station's infrastructure. Operating at a rated voltage of 6600V and a rated current of 125A, these pumps, totaling five in number, collectively boast a remarkable capacity, capable of pumping an impressive 590 million liters of water per day. Notably, these pumps are equipped with an array of sensors meticulously installed to gather crucial operational data. These sensors capture various parameters including vibration readings, temperature measurements of the pump, as well as voltage and current readings, enabling comprehensive monitoring and maintenance strategies. This section endeavors to delve into the intricacies of these sensors, elucidating their role in ensuring optimal performance and longevity of the centrifugal pumps, thereby fortifying the reliability of the water supply systems managed by the BWSSB.

Figure 3.1 Pumps installed in TK Halli

3.3 Sensors Used for Monitoring

Embedded within the centrifugal pump system at the TK Halli pumping station administered by the Bangalore Water Supply and Sewerage Board (BWSSB) are sophisticated sensor arrays meticulously engineered to bolster operational oversight and maintenance efficacy. Vibration sensors, strategically positioned, diligently capture four-point vibration data, intricately monitoring the pump's drive end and non-drive end, as well as the motor's drive end and non-drive end. This comprehensive vibration analysis enables early detection of potential mechanical issues, facilitating predictive maintenance interventions to forestall operational disruptions.

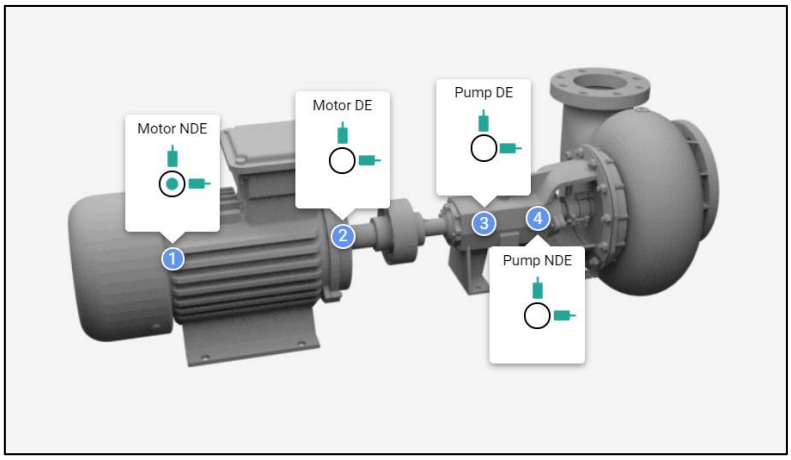


Figure 3.2 Vibration measurement points

Figure 3.3 Vibration sensors installed on pumps

In tandem with vibration monitoring, voltage and current sensors are seamlessly integrated into the system, meticulously tracking the motor's electrical consumption. These sensors provide real-time insights into power usage patterns, enabling operators to optimize energy efficiency and mitigate the risk of electrical anomalies, thus enhancing overall system reliability and cost-effectiveness.

Furthermore, temperature sensors play a pivotal role in safeguarding pump and motor integrity by furnishing precise four-point temperature readings. By monitoring temperature variations at critical junctures, including the pump's drive and non-drive ends, as well as the motor's drive end and non-drive end, potential overheating issues can be promptly identified and addressed, averting costly equipment failures and prolonging asset lifespan.

This comprehensive sensor integration paradigm empowers BWSSB personnel with actionable intelligence, facilitating data-driven decision-making and proactive maintenance initiatives. By leveraging advanced sensor technologies, the centrifugal pump system at TK Halli pumping station stands poised to deliver sustained performance, bolstering the resilience and efficiency of the water supply infrastructure serving the community.

3.4 Dataflow Diagram

A Dataflow Diagram (DFD) provides a graphical representation of how data moves through a system. It illustrates the flow of data from input to processing and finally to output, helping to visualize the steps involved in handling and transforming the data. In the pump anomaly detection system, the DFD outlines the process from data collection through sensors, feature extraction, and the application of machine learning models to classify pump states. This diagram helps stakeholders understand the system's workings and ensures that all necessary data handling steps are included and properly sequenced. Additionally, a well-constructed DFD highlights potential bottlenecks and areas for improvement, facilitating the optimization of the data processing workflow. It also serves as a valuable documentation tool, aiding in communication and coordination among development and maintenance teams.

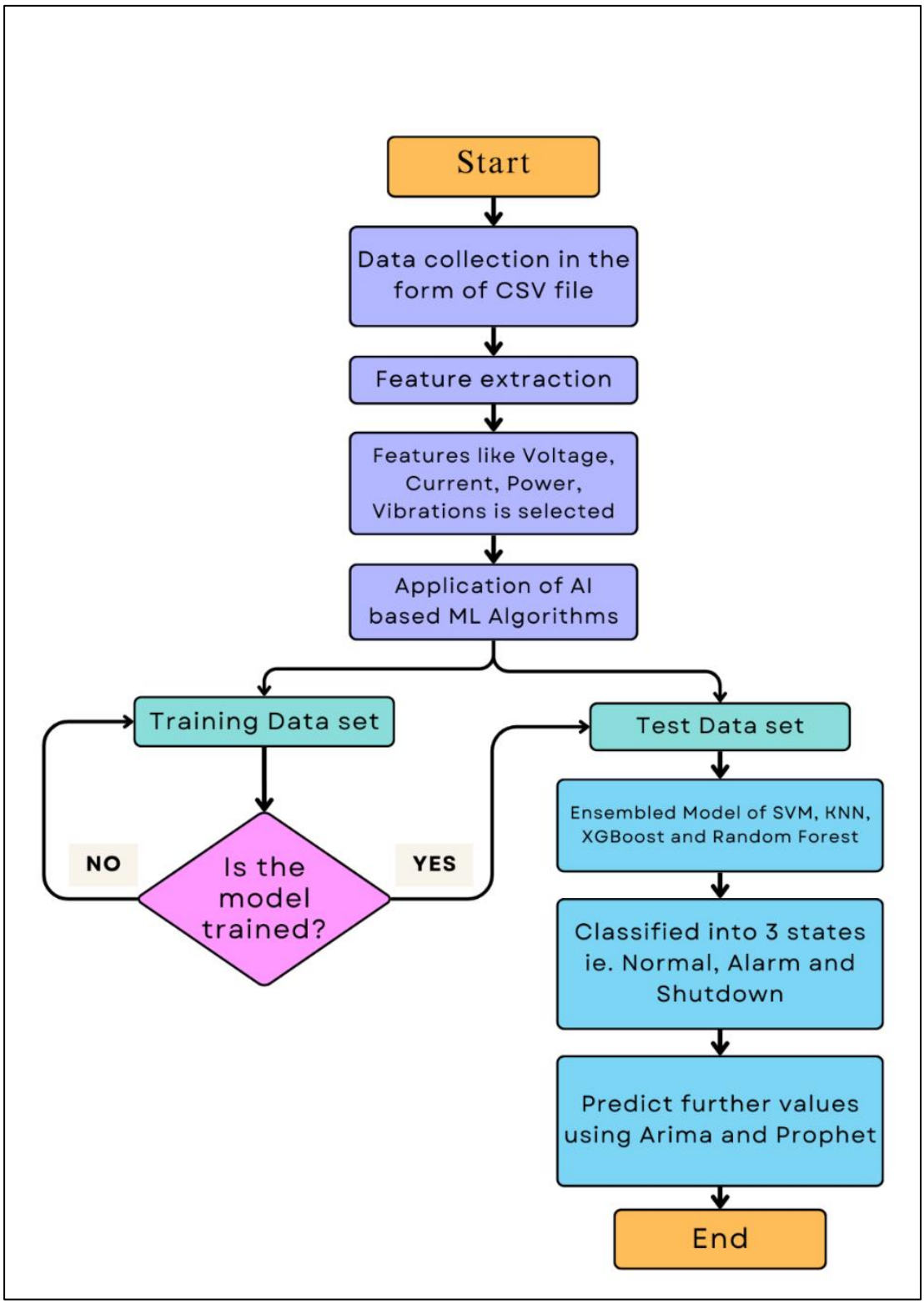


Figure 3.4 Dataflow Diagram

The Data Flow Diagram (DFD) serves as a comprehensive visual representation of the predictive maintenance process for pump systems, delineating the sequential flow of data from its collection to the final classification outcome. Initially, the process commences with

the collection of raw data, typically in the form of a CSV file. This raw data serves as the foundation for subsequent stages of analysis. Following data collection, the next step involves feature extraction. Relevant features such as Voltage, Current, Power, Temperature, and Vibrations are extracted from the raw data. This stage is crucial as it identifies the key parameters that are indicative of the pump system's performance and health. Subsequently, the extracted features undergo analysis through AI-based Machine Learning (ML) algorithms. These algorithms are instrumental in training a predictive model using a designated training dataset. The model learns patterns and relationships within the data, enabling it to make predictions based on new input. To evaluate the efficacy of the trained model, a separate test dataset is employed for testing purposes. This step assesses the model's performance and its ability to generalize to unseen data accurately. The final stage of the process involves ensembling the predictions from multiple ML algorithms. Various techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost are employed to classify pump states into distinct categories, namely Normal, Alarm, and Shutdown. By integrating predictions from diverse algorithms, the ensemble approach enhances the robustness and reliability of the classification outcome, thereby facilitating effective decision-making regarding maintenance interventions for pump systems.

3.5 Hardware Requirements

- Intel i3/i5 2.4 GHz processor
- 500 GB hard drive
- 4/8 GB RAM

3.6 Software Requirements

- Operating System: Windows 8 or Higher
- Programming Language: Python 3.6 (or above) & related libraries
- Visualisation Tools: Power BI, Microsoft SSMS

3.6.1 Power BI

Power BI is a business analytics tool developed by Microsoft that enables users to visualize and analyze data. It offers a user-friendly interface with drag-and-drop functionality, allowing users to create interactive reports and dashboards without requiring extensive technical expertise. Power BI connects to a variety of data sources, including databases, spreadsheets, and cloud services, enabling users to import and transform data easily. With its robust data modeling capabilities, Power BI allows users to create relationships between different data tables and perform complex calculations. Additionally, Power BI offers powerful visualization options, including charts, graphs, maps, and gauges, to present data insights effectively. Users can also collaborate and share reports securely within their organization or publish them to the web. Overall, Power BI empowers businesses to gain valuable insights from their data and make informed decisions.

3.6.2 SQL Server and Server management studio

SQL Server, a Microsoft RDBMS, manages structured data with storage, retrieval, and security features. SQL Server Management Studio (SSMS), its IDE, aids database management, query execution, and schema design. SSMS integrates tools like Query Editor for SQL execution, Object Explorer for database navigation, and Visual Database Designer for schema visualization. Performance monitoring tools help analyze server activity, while import/export wizards facilitate data movement. Security management features ensure user authentication and permissions. Together, SQL Server and SSMS offer a robust platform for database development, administration, and management, empowering businesses to efficiently handle their data operations.

3.6.3 Google Collab

Google Colab, a cloud-based platform by Google, offers Python coding in browser. Integrated with Jupyter Notebooks, it provides free GPU/TPU access, ideal for machine learning. Its cloud execution eliminates local installation needs, enhancing accessibility. Collaboration features allow real-time sharing and teamwork. Seamlessly linked with Google Drive, managing projects becomes effortless. Pre-installed with popular libraries like TensorFlow, PyTorch, and Scikit-learn, it supports diverse tasks. With its convenience, power, and extensive features, Google Colab emerges as a favored environment for Python coding, especially in data analysis and machine learning projects.

3.6.4 VS code

Visual Studio Code (VS Code) is a versatile and feature-rich code editor developed by Microsoft. It offers a user-friendly interface with a wide range of functionalities tailored for software development across various programming languages. With its extensive library of extensions, VS Code can be customized to suit individual preferences and workflows. It provides built-in support for version control systems like Git, making collaboration and code management seamless. Moreover, its integrated terminal allows developers to execute commands and run scripts directly within the editor. VS Code's powerful debugging capabilities, intelligent code completion, and extensive language support make it a popular choice among developers for building and debugging applications efficiently.

3.6.5 Python

Python has emerged as the preferred language for machine learning (ML) due to its simplicity, versatility, and extensive library support. With a rich ecosystem that includes TensorFlow, PyTorch, Scikit-learn, and Keras, Python offers a comprehensive toolkit for ML development. Its ease of learning, facilitated by clear and concise syntax, makes it accessible to both beginners and experts. Moreover, Python's flexibility allows developers to implement ML algorithms efficiently, supporting various programming paradigms. Supported by a large and active community, Python benefits from extensive documentation, tutorials, and online resources, aiding in learning and problem-solving. Integration with other languages and tools further enhances its utility for tasks such as data preprocessing, visualization, and deployment. While Python may not match the performance of lower-level languages, its scalability through extensions and optimizations, along with GPU acceleration in libraries like TensorFlow and PyTorch, ensures efficient computation for demanding tasks.

3.6.6 Scikit-Learn

Scikit-learn, a prominent Python library for machine learning, offers a versatile toolkit for various tasks including classification, regression, clustering, and dimensionality reduction. Known for its user-friendly interface, it facilitates implementation of algorithms with minimal code. With extensive documentation and tutorials, it is accessible to both beginners and experts. Scikit-learn integrates seamlessly with other Python libraries such as NumPy,

Pandas, and Matplotlib, enhancing its capabilities. Its robustness, efficiency, and broad range of functionalities make it a go-to choose for many machine learning practitioners and researchers.

3.6.7 NumPy

NumPy is a fundamental library in Python used for numerical computing. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy's array operations are optimized for performance, making it a cornerstone of many scientific and engineering applications. It serves as the foundation for many other Python libraries in the data science and machine learning ecosystem, enabling high-performance computation and data manipulation.

3.6.8 Pandas

Pandas is a popular Python library tailored for data manipulation and analysis. It introduces two primary data structures: Series (1-dimensional) and DataFrame (2-dimensional), enabling intuitive handling of tabular data. Pandas excels at tasks like data cleaning, exploration, and transformation, offering a plethora of functions and methods for these operations. With its powerful indexing and selection capabilities, it simplifies data slicing and filtering. Additionally, Pandas integrates seamlessly with other libraries, facilitating interoperability within the Python ecosystem. Its versatility and user-friendly interface make it indispensable for data professionals, from data scientists to analysts, in streamlining their workflows and extracting insights from datasets.

3.6.9 Prophet

Prophet, developed by Facebook, is a valuable tool for time series forecasting in machine learning. Its key contribution lies in its ability to handle time series data with seasonal patterns, holidays, and other effects effortlessly. Prophet simplifies the process of building accurate and interpretable time series models, making it accessible to users with varying levels of expertise. Its intuitive interface and robust performance make it an essential tool for ML practitioners seeking to generate reliable forecasts from their data while accounting for complex temporal dynamics.

3.6.10 StatsModel

StatsModels is a Python module for statistical modeling and hypothesis testing. It offers functions to estimate various models like linear regression, GLMs, and mixed-effects models. Additionally, it provides tools for time series analysis, including ARIMA and VAR models. StatsModels supports hypothesis testing, with a range of statistical tests available. It integrates with Matplotlib and Seaborn for data visualization. Flexible and extensible, StatsModels is widely used in statistics, data science, and research for its comprehensive capabilities in analyzing data and building predictive models.

3.7 Functional Requirements

- **Data Acquisition:** Retrieve pump system data from various sources, including sensors, IoT devices, and historical databases, ensuring compatibility with the Pump Academy dataset.
- **Preprocessing:** Perform comprehensive data preprocessing to handle missing values, outliers, and noise, ensuring data quality and consistency for downstream analysis.
- **Feature Engineering:** Conduct feature engineering to extract relevant features from the pump system data, including voltage, power, current, and other relevant variables, to capture patterns indicative of pump health and performance.
- **Machine Learning Models:** Implement machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and XGBoost, to classify pump conditions into normal, alarm, and shutdown states based on the extracted features.
- **Evaluation Metrics:** Utilize evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the classification models and aid in model selection.
- **Forecasting Models:** Implement forecasting models, such as Prophet, Vector Auto Regressive (VAR), and Autoregressive Integrated Moving Average (ARIMA), to predict future pump conditions and anticipate potential failures or maintenance needs.
- **Visualization:** Provide visualizations of model predictions, including classification results and forecasted pump conditions, through interactive dashboards or graphical representations to facilitate interpretation and decision-making.

3.8 Non-Functional Requirements

- **Performance:** The system must efficiently process and analyse pump data to detect anomalies promptly, even with large datasets or high-resolution sensor readings.
- **Scalability:** As the number of pumps and data points grows, the system should scale seamlessly to handle the increased load without compromising performance.
- **Reliability:** Anomaly detection algorithms must be robust to variations in pump behaviour and environmental factors, minimizing false positives and ensuring accurate detection of anomalies.
- **Availability:** The system should operate continuously, with minimal downtime, ensuring uninterrupted monitoring and timely detection of pump anomalies.
- **Usability:** The user interface should be intuitive, allowing operators to easily configure monitoring parameters, visualize pump performance, and interpret anomaly alerts without extensive training.
- **Maintainability:** Design the system with modular components and well-documented code to facilitate maintenance and updates. This includes providing clear guidelines for troubleshooting, version control practices, and documentation for future enhancements or modifications.
- **Compatibility:** Ensure that the system is compatible with a wide range of pump systems, sensors, and data formats commonly used in industrial settings. This includes supporting industry-standard communication protocols and data exchange formats to enable seamless integration with existing infrastructure.
- **Interoperability:** Ensure that the system can seamlessly exchange data and communicate with other enterprise systems, such as asset management systems, enterprise resource planning (ERP) systems, and maintenance management systems. This interoperability enables seamless data flow and integration across different organizational functions.

CHAPTER 4
SYSTEM DESIGN

CHAPTER 4

SYSTEM DESIGN

4.1 Architecture Diagram

The architecture diagram presents the high-level structure of the system, showing the major components and their interactions. It helps in understanding how the system is organized and how different parts communicate with each other. For the pump anomaly detection system, the architecture diagram includes components such as data acquisition from sensors, data preprocessing, machine learning model training and classification, and data storage and visualization. This diagram is essential for developers and engineers as it provides a blueprint for implementing and integrating the system components effectively.

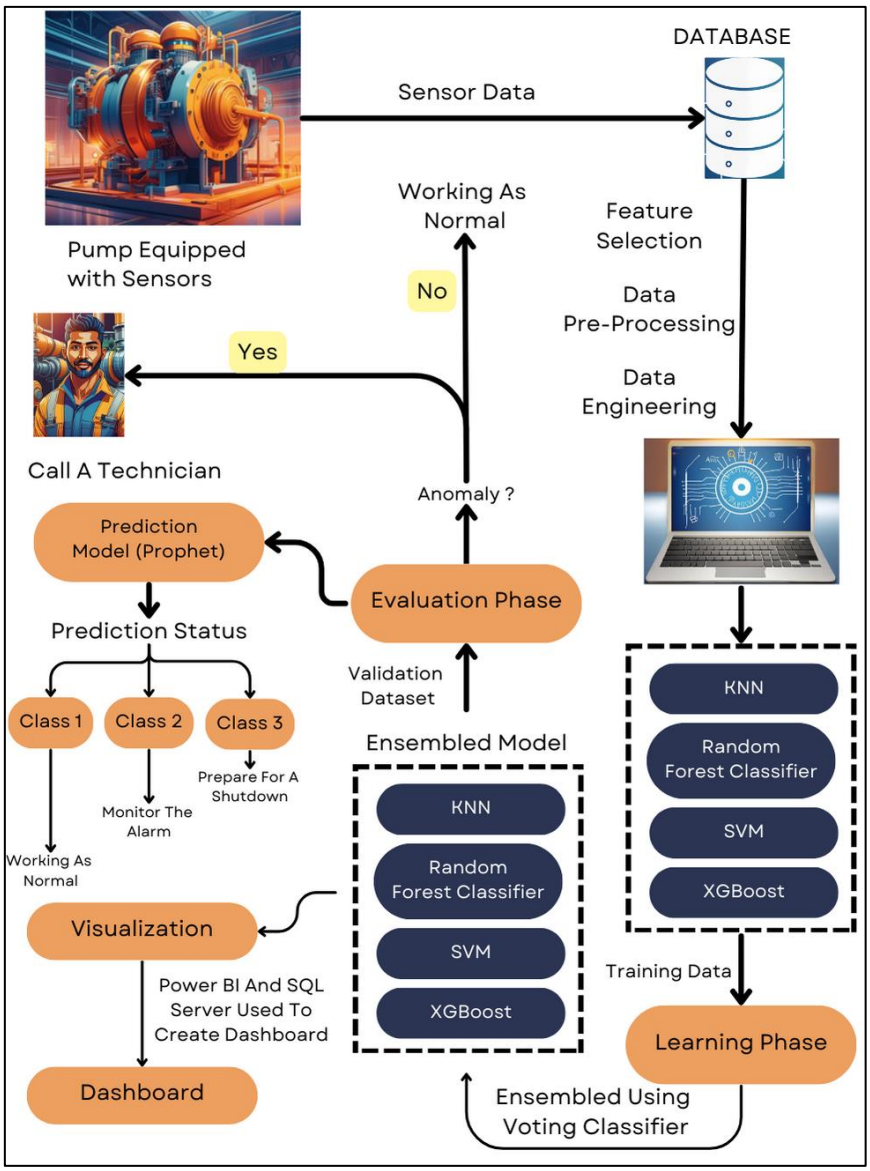


Figure 4.1 Architecture Diagram

The process commences with the acquisition of pump dataset through sensors installed on the pump, which capture crucial operational parameters. These sensors serve as the primary data source, providing real-time information about the pump's performance. Following data acquisition, the dataset undergoes feature engineering and data preprocessing techniques to refine and enhance its quality. This stage involves extracting meaningful features from the raw data and applying necessary transformations to prepare it for analysis. By optimizing the dataset in this manner, it becomes more conducive to accurate and insightful analysis.

Next, utilizing trained machine learning algorithms, a classifier is developed to categorize the dataset into distinct classes. These classes typically represent different operational states or conditions of the pump, such as normal operation, maintenance required, or malfunction. By employing advanced classification techniques, the classifier facilitates efficient analysis and decision-making, enabling stakeholders to respond promptly to any deviations from optimal performance. Once the dataset has been classified, the categorized data is stored and managed in a SQL Server database. This database serves as a centralized repository, ensuring data integrity, security, and accessibility. By storing the data in a structured format within a relational database management system (RDBMS), stakeholders can easily retrieve and query the data as needed, supporting various analytical and reporting requirements.

To provide stakeholders with real-time insights into pump performance and operational status, the classified data is retrieved from the database and visualized using PowerBI. PowerBI is a powerful business intelligence tool that enables the creation of dynamic dashboards and interactive visualizations. By leveraging PowerBI's capabilities, stakeholders gain access to intuitive and user-friendly dashboards that present key performance indicators (KPIs), trends, and anomalies in pump operations. This visualization empowers stakeholders to monitor pump performance proactively and make informed decisions to optimize operational efficiency and minimize downtime.

4.2 Use case Diagram

A use case diagram is a visual representation that outlines the interactions between users (actors) and a system to achieve specific goals. It helps in identifying the functional requirements of a system and clarifies the relationships between different components and stakeholders. The advantages of use case diagrams include providing a clear and concise overview of system functionalities, facilitating communication among team members, ensuring all user interactions are considered, and aiding in the design and development process. Additionally, they help in identifying potential issues early in the project lifecycle and serve as a useful reference for documentation and future maintenance.

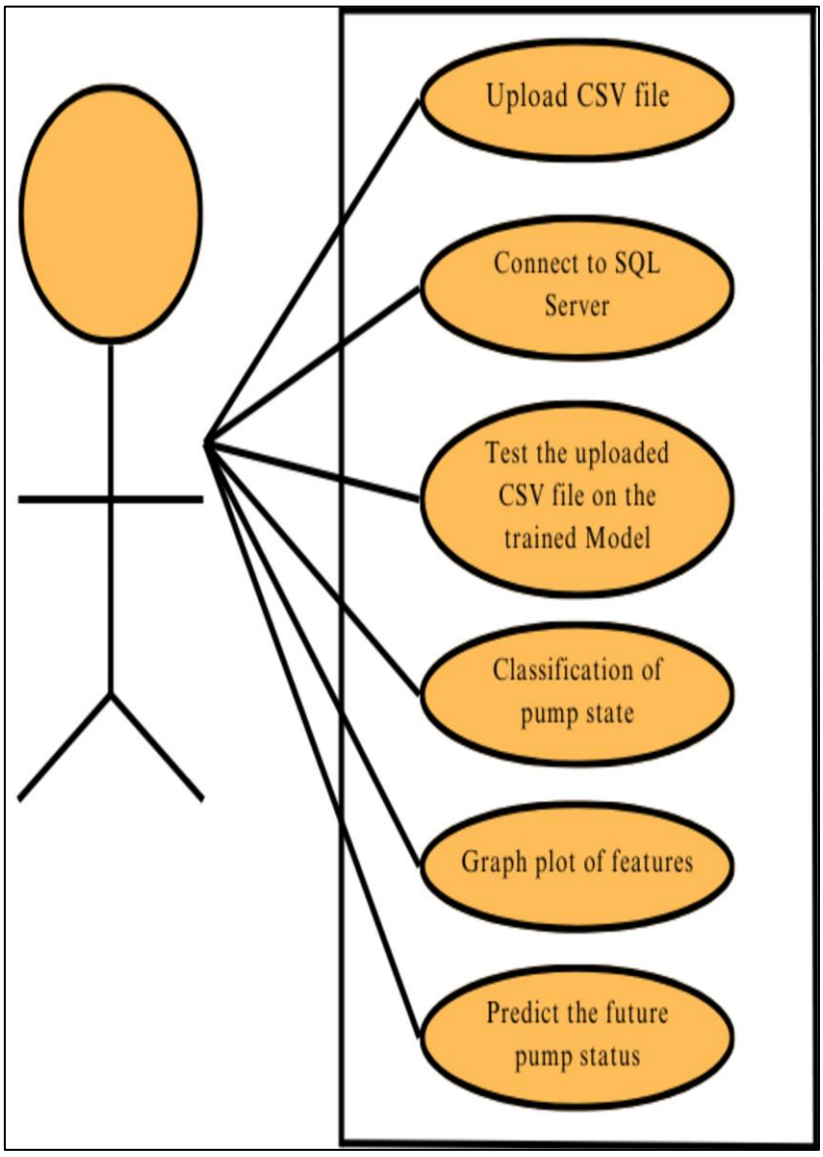


Figure 4.2 Use case Diagram

- **Upload Pump Data (User):** The process begins with the user uploading a CSV file containing comprehensive pump data acquired from sensors. This data typically includes parameters such as voltage, current, power, temperature, and vibrations, capturing vital operational insights.
- **Connect to SQL Server:** Subsequently, the uploaded data establishes a connection with an SQL Server database, ensuring robust data management and storage capabilities. This connection facilitates efficient organization, retrieval, and manipulation of the pump data, ensuring data integrity and accessibility throughout the analysis process.
- **Test Data on Pre-Trained ML Model:** The uploaded pump data undergoes testing on a pre-trained machine learning model specifically designed for pump state classification. This model has been trained using historical pump data and employs sophisticated algorithms to classify the incoming data into distinct operational states, such as normal operation, alarm, or shutdown. Through this classification process, the system gains valuable insights into the current status and condition of the pump.
- **Classification of Pump Data:** Leveraging the insights from the pre-trained machine learning model, the pump data is classified into relevant categories based on its present characteristics and historical trends. This classification enables stakeholders to identify any deviations or anomalies in pump performance, facilitating proactive maintenance interventions and decision-making.
- **Further Analysis and Visualization:** Following classification, the classified pump data undergoes further analysis to extract additional insights and identify underlying patterns or trends. This analysis may include statistical techniques, trend analysis, or anomaly detection algorithms to uncover valuable information about the pump's operational behavior. Additionally, the data is visualized through graphs, charts, and dashboards to provide stakeholders with intuitive and actionable insights into pump performance and health.
- **Prediction of Future Pump Status:** Utilizing the insights gained from the analysis, the system employs predictive modeling techniques to forecast the future status of the pump. By extrapolating trends and patterns identified in the data, the system predicts potential performance trajectories and maintenance requirements, enabling stakeholders to proactively address issues and optimize pump operations.

CHAPTER 5
IMPLEMENTATION

CHAPTER 5

IMPLEMENTATION

5.1 EDA (Exploratory Data Analysis)

In the exploratory data analysis (EDA) phase, we conducted a comprehensive examination of the pump system data by plotting voltage, current, and power against time. This allowed us to visualize the temporal patterns and relationships between these variables. Furthermore, KNN clustering was applied to the power data to identify distinct clusters based on their characteristics.

Figure 5.1 Current Vs Time

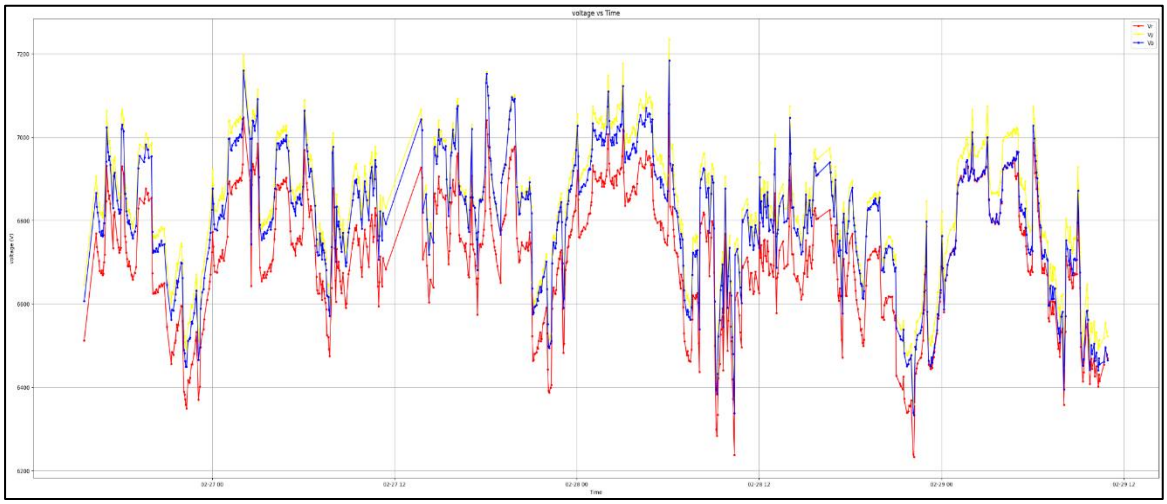


Figure 5.2 Voltage Vs Time

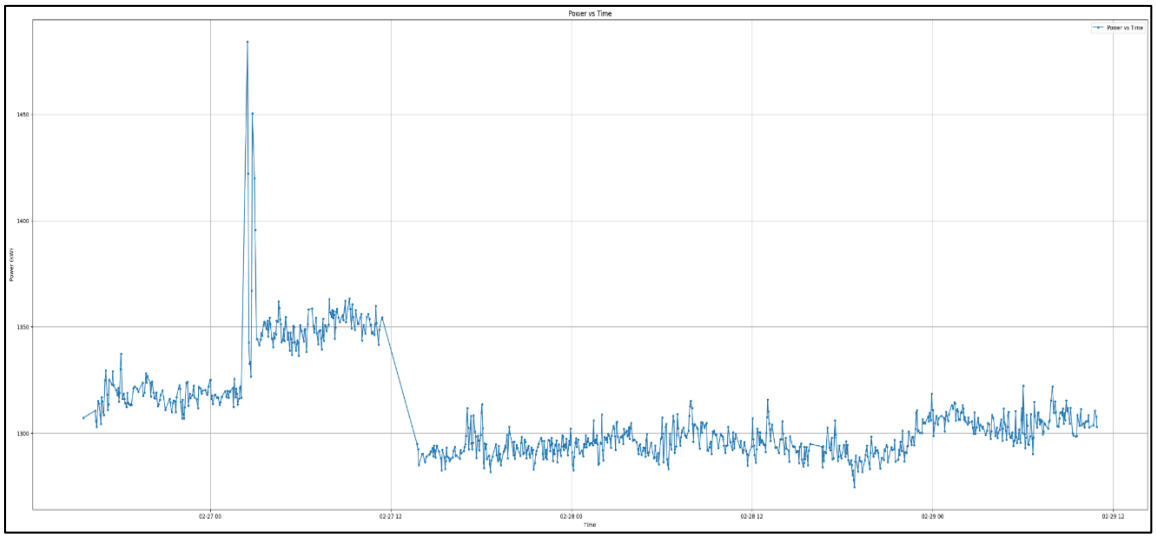


Figure 5.3 Power Vs Time

From the plots generated during EDA, it was observed that an increase in power consumption tends to coincide with an increase in current, indicating a potential relationship between these variables. This finding provides valuable insight into the behaviour of the pump system under different operating conditions.

Moreover, the application of K-Means clustering facilitated the categorization of pump power, voltage, and current into three distinct states: normal, alarm, and shutdown. By clustering the power data, we were able to identify patterns indicative of different operational states of the pump system.

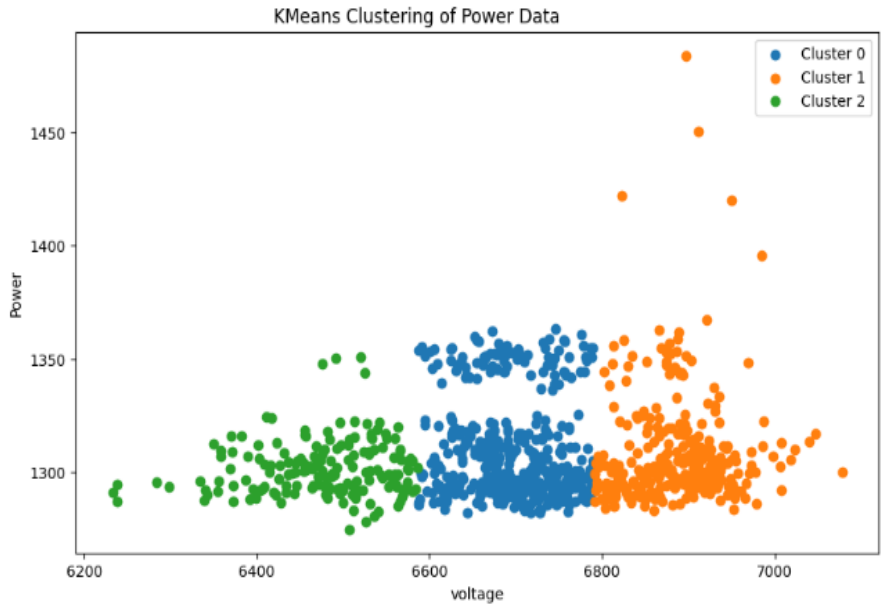


Figure 5.4 K-Means clustering of power data wrt voltage value

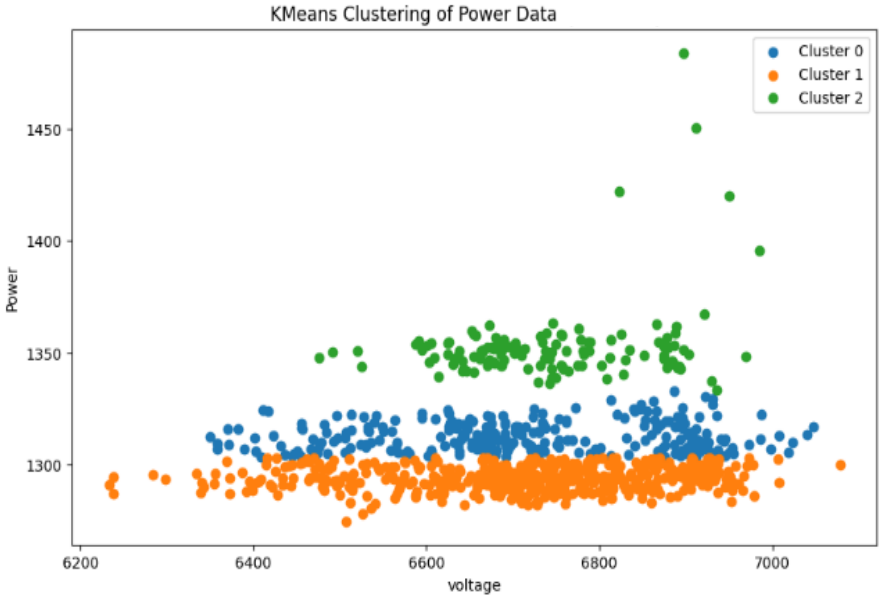


Figure 5.5 K-Means clustering of power data wrt voltage value

The image 5.4 and 5.5 depict the results of K-Means clustering applied to power data concerning voltage values, enabling the categorization of pump operational states into normal, alarm, and shutdown. Through this analysis, distinct patterns indicative of each state are identified, providing valuable insights into the performance and behavior of the pump system.

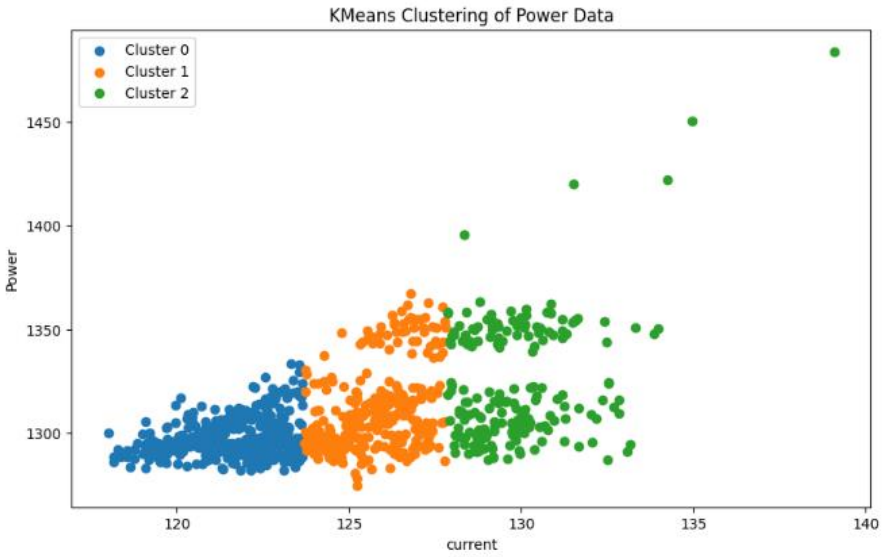


Figure 5.6 K-Means clustering of power data wrt current value

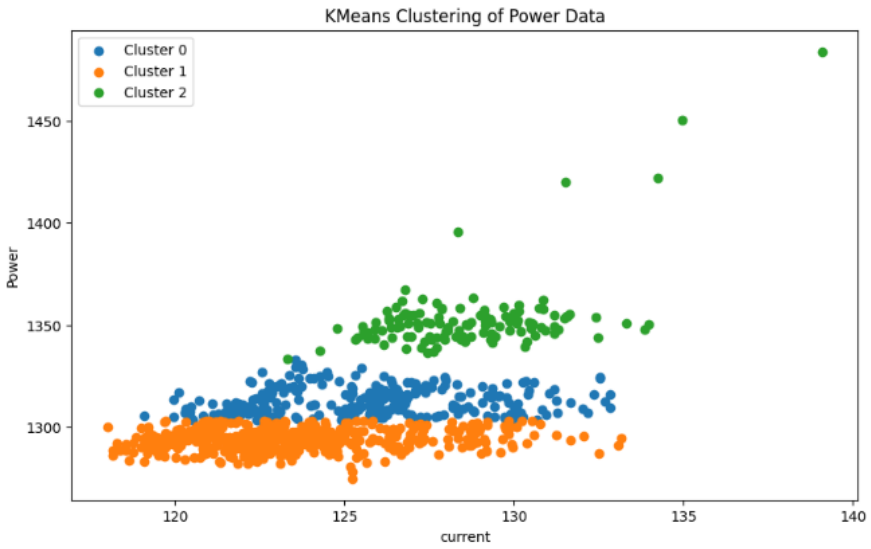


Figure 5.7 K-Means clustering of power data wrt current value

The image 5.6 and 5.7 depict the results of K-Means clustering applied to power data concerning current values, enabling the categorization of pump operational states into normal, alarm, and shutdown. Through this analysis, distinct patterns indicative of each state is identified, providing valuable insights into the performance and behavior of the pump system.

5.2 Labelling Dataset

The dataset underwent a labelling process wherein observations were categorized into three distinct classes:

- Class 0: Representing normal operating conditions.
- Class 1: Representing instances triggering an alarm state.
- Class 2: Indicating instances necessitating a system shutdown.

These classifications were determined based on predetermined threshold values established for each of the following parameters as shown in table below

SL NO	Parameter	Base value	Normal (0)	Alarm (1)	Shutdown (2)
1	Voltage	6600 V	$< \pm 6\%$	$\geq \pm 6\%$ & $< \pm 8\%$	$\geq \pm 8\%$
2	Current	125 A	$< \pm 3.5\%$	$\geq \pm 3.5\%$ & $< \pm 6\%$	$\geq \pm 6\%$
3	Power	1250 KV	$< \pm 3.5\%$	$\geq \pm 3.5\%$ & $< \pm 6\%$	$\geq \pm 6\%$
4	Vibration (mm/s)	-	< 4.2	≥ 4.2 & < 7.2	≥ 7.2
5	Temperature (°C)	-	<49	≥ 49 & <54	≥ 54

Table 5.1 Threshold values of various parameters

5.3 Application of Machine Learning Models

In our analysis, we utilized a diverse array of machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost. These models were separately trained and evaluated for each parameter: power consumption, vibration levels, and temperature readings.

5.3.1 KNN (K-Nearest Neighbors)

We utilized the K-Nearest Neighbors (KNN) algorithm to classify the dataset based on various parameters such as power consumption, vibration levels, and temperature readings. KNN is a straightforward yet powerful classification algorithm that assigns class labels to unclassified data points based on the labels of its nearest neighbors in the feature space. We first pre-processed the dataset by standardizing the features to remove the mean and scale to unit variance, ensuring uniformity in the data distribution. Subsequently, we split the data into training and testing sets using the train_test_split function from the sklearn library, with a test size of 20%. For each parameter of interest, namely vibration levels, temperature readings, and power data (voltage and current), we instantiated a KNN classifier with a chosen value of k (number of neighbors) and trained it on the training data. The classifiers were then used to make predictions on the testing data to evaluate their performance.

The evaluation metrics used to assess the classifiers' performance included confusion matrices and classification reports, which provided insights into the models' accuracy, precision, recall, and F1-score. By analyzing these metrics, we gained valuable insights into the effectiveness of the KNN algorithm in classifying the dataset based on different parameters.

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1465
1	0.97	0.98	0.97	526
2	1.00	0.86	0.93	22
accuracy			0.99	2013
macro avg	0.99	0.94	0.96	2013
weighted avg	0.99	0.99	0.99	2013

Figure 5.8 Classification Report for KNN

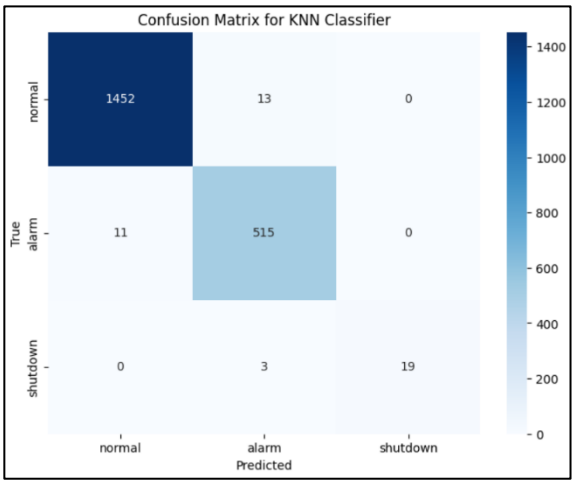


Figure 5.9 Confusion Matrix for KNN

5.3.2 SVM (Support Vector Machine)

We incorporated the Support Vector Machine (SVM) algorithm to classify the dataset based on power consumption, vibration levels, and temperature readings. SVM is a robust supervised learning algorithm known for its effectiveness in handling classification tasks, particularly in scenarios with high-dimensional data and non-linear decision boundaries. We first pre-processed the dataset by standardizing the features using the StandardScaler to remove the mean and scale to unit variance. Subsequently, we split the data into training and testing sets using the train_test_split function from sklearn, with a test size of 20%. For the classification task, we instantiated an SVM classifier with a polynomial kernel ('poly') to capture complex relationships in the data. The classifier was then trained on the training data to learn the underlying patterns in the dataset.

Following the training phase, we utilized the trained SVM model to make predictions on the testing data. The predictions were evaluated using various metrics, including a confusion matrix and a classification report, which provided insights into the model's accuracy, precision, recall, and F1-score.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1484
1	0.99	0.99	0.99	520
2	1.00	1.00	1.00	9
accuracy			0.99	2013
macro avg	1.00	1.00	1.00	2013
weighted avg	0.99	0.99	0.99	2013

Figure 5.10 Classification Report for SVM

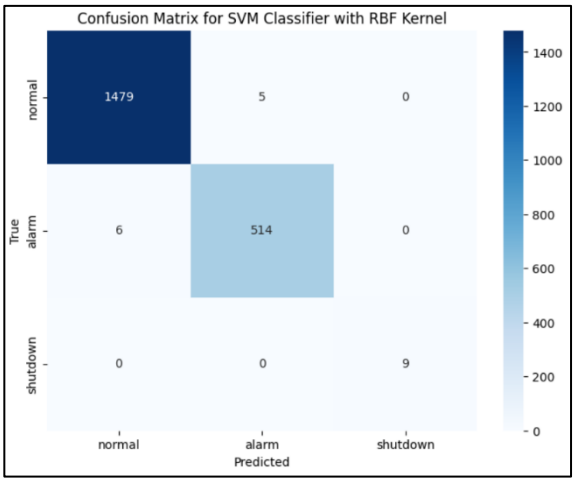


Figure 5.11 Confusion Matrix for SVM

5.3.3 XGBoost

We incorporated the XGBoost classifier, a popular gradient boosting algorithm, to classify the dataset based on power consumption, vibration levels, and temperature readings. XGBoost is known for its efficiency, scalability, and high performance, making it a suitable choice for our analysis of pump system data. After reading the preprocessed dataset from a CSV file, we preprocessed the data by standardizing the features using the StandardScaler to ensure uniform scaling across all variables. The dataset was then split into training and testing sets using the train_test_split function from sklearn, with a test size of 20%. Next, we instantiated an XGBoost classifier to leverage its gradient boosting capabilities for improved classification accuracy. XGBoost is particularly effective in handling complex datasets and capturing intricate patterns in the data. We then trained the XGBoost classifier on the training data to learn the underlying patterns and relationships between the features and target variable. The trained model was used to make predictions on the testing data to evaluate its performance.

To assess the classifier's performance, we calculated various metrics including a confusion matrix and a classification report. These metrics provided insights into the model's accuracy, precision, recall, and F1-score, allowing us to evaluate its effectiveness in classifying pump system data accurately.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1480
1	1.00	0.99	1.00	521
2	1.00	1.00	1.00	12
accuracy			1.00	2013
macro avg	1.00	1.00	1.00	2013
weighted avg	1.00	1.00	1.00	2013

Figure 5.12 Classification Report for XGBoost

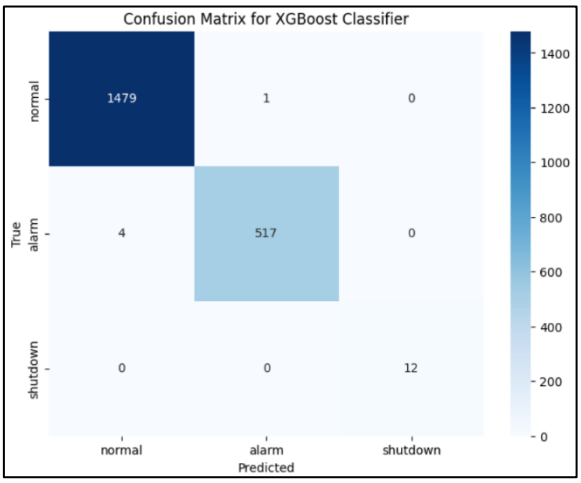


Figure 5.13 Confusion Matrix for XGBoost

5.3.4 Random Forest

We integrated the Random Forest classifier, a powerful ensemble learning algorithm, to classify the dataset based on power consumption, vibration levels, and temperature readings. Random Forest is known for its robustness and ability to handle both classification and regression tasks effectively, making it well-suited for our analysis of pump system data. After reading the preprocessed dataset from a CSV file, we preprocessed the data by standardizing the features using the StandardScaler to ensure uniform scaling across all variables. The dataset was then split into training and testing sets using the train_test_split function from sklearn, with a test size of 20%. Next, we instantiated a Random Forest classifier with 90 decision trees (n_estimators=90) to harness the power of ensemble learning. Ensemble methods like Random Forest combine multiple base estimators to improve generalizability and reduce overfitting. Subsequently, we trained the Random Forest classifier on the training data to learn the underlying patterns and

relationships between the features and target variable. The trained model was then used to make predictions on the testing data to evaluate its performance.

We assessed the classifier's performance using various metrics, including a confusion matrix and a classification report. These metrics provided insights into the model's accuracy, precision, recall, and F1-score, allowing us to gauge its effectiveness in classifying pump system data accurately.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1478
1	1.00	1.00	1.00	520
2	0.94	1.00	0.97	15
accuracy			1.00	2013
macro avg	0.98	1.00	0.99	2013
weighted avg	1.00	1.00	1.00	2013

Figure 5.14 Classification Report for Random Forest

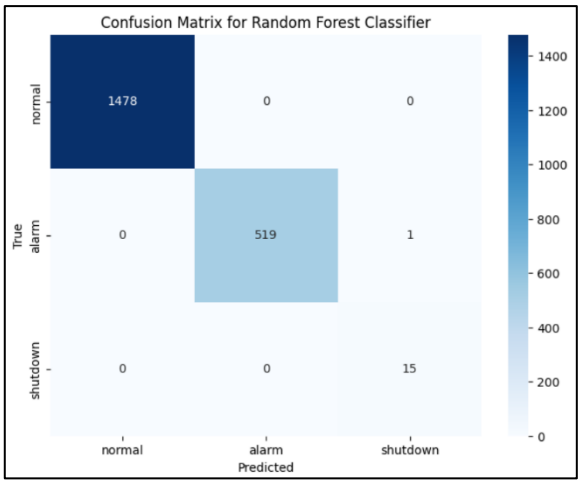


Figure 5.15 Confusion Matrix for Random Forest

SL Num	Methods	Performances	Limitation
1	KNN	• Accuracy: 99% • Recall: 94% • Precision: 99%	low with Big Data, Sensitive to Features, Requires Feature Scaling, Struggles with Many Features, Choosing the Right k
2	SVM	• Accuracy: 99% • Recall: 99% • Precision: 99%	Inefficient with Large Datasets, Prone to Overfitting if Not Properly Tuned, Sensitivity to Noise in the Data
3	Random Forest	• Accuracy: 100% • Recall: 100% • Precision: 100%	Susceptible to Overfitting, especially with noisy data
4	LSTM	• Accuracy: 90% • Recall: 79% • Precision: 80%	Less effective with short sequences, Complex architecture may lead to longer training times
5	XGBoost	• Accuracy: 99% • Recall: 97% • Precision: 98%	Sensitive to overfitting if hyperparameters are not tuned properly
6	conv1D and LSTM	• Accuracy: 70% • Recall: 68% • Precision: 58%	Limited by short-term memory, Longer training times due to complex architecture
7	Random Forest and LSTM	• Accuracy: 88% • Recall: 85% • Precision: 86%	Susceptible to overfitting, Longer training times due to combining multiple models

Table 5.2 summary of all ML models

5.3.5 Ensembled Model

To further enhance the predictive capabilities and overall accuracy of our system, we adopted an ensemble approach, specifically the voting method. By combining the predictions of multiple base models, we aimed to mitigate individual model biases and leverage the collective intelligence of the ensemble to make more robust and reliable predictions. The ensemble of models was trained on the same dataset and evaluated using the same metrics as the individual models. Through this ensemble method, we aimed to capitalize on the strengths of each base model while mitigating their weaknesses, ultimately striving for improved accuracy and predictive performance in classifying system states.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1478
1	1.00	1.00	1.00	520
2	0.94	1.00	0.97	15
accuracy			1.00	2013
macro avg	0.98	1.00	0.99	2013
weighted avg	1.00	1.00	1.00	2013

Figure 5.16 Classification Report for Ensembled Model

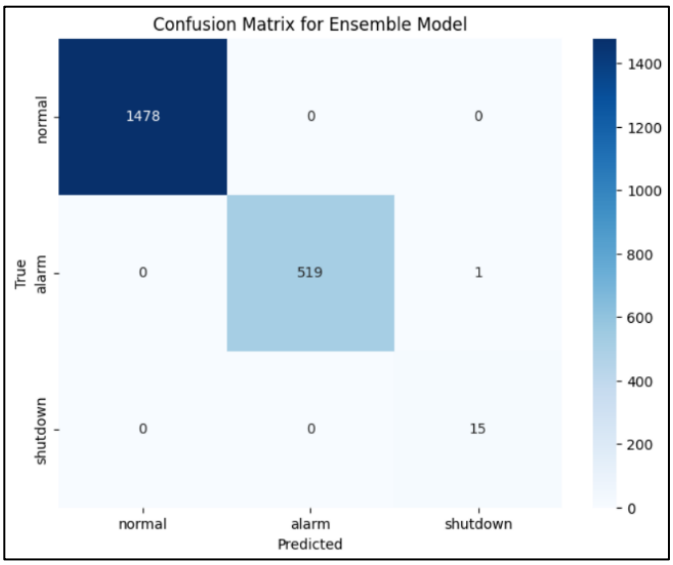


Figure 5.17 Confusion Matrix for Ensembled Model

5.4 Predictive Maintenance through Forecasting

In our pursuit of predictive maintenance, we extended our analysis to forecast future values of power consumption, temperature, and vibration levels. To achieve this, we employed sophisticated forecasting methods including AutoRegressive Integrated Moving Average (ARIMA) and Prophet. While Prophet offers a versatile interface with robust capabilities for time series forecasting, ARIMA, a classical statistical approach, provides insights into power consumption trends with its rigorous methodology and statistical foundations. These complementary techniques enhance our predictive maintenance strategies by providing accurate forecasts of equipment behavior.

5.4.1 ARIMA (AutoRegressive Integrated Moving Average)

ARIMA (AutoRegressive Integrated Moving Average) model for time series forecasting, specifically focusing on predicting power consumption trends based on historical data. The implementation begins with the loading and preprocessing of the dataset, which involves converting the 'Date' column to datetime format and handling any missing values using forward filling techniques to ensure data integrity. Subsequently, the dataset is split into training and testing subsets, with 80% of the data allocated for training the model and the remaining 20% for evaluating its performance. The core of the implementation revolves around fitting the ARIMA model to the training data. Here, an ARIMA model is instantiated with a defined order parameter (p, d, q), which signifies the number of autoregressive, differencing, and moving average terms, respectively. In this scenario, the order is set to (5, 0, 2), although fine-tuning of these parameters may be necessary to optimize model performance based on the dataset's characteristics. Once instantiated, the model is trained on the training data using the fit method. Following model training, forecasts are generated for the test set using the trained ARIMA model. The forecast method is employed to predict power consumption values for the duration of the test set. These forecasted values are then evaluated against the actual values from the test set to quantify the model's performance. This evaluation is conducted using the Root Mean Squared Error (RMSE) metric, which measures the deviation between the predicted and actual values. A lower RMSE value indicates a closer alignment between the predicted and observed values, reflecting greater forecasting accuracy.

Finally, the results of the ARIMA model, including the training data, test data, and forecasted values, are visualized using Matplotlib. This visualization aids in assessing the model's predictive capabilities by providing a graphical representation of how well the forecasted values align with the actual power consumption trends observed in the test data. Overall, this implementation demonstrates the practical application of ARIMA modeling in forecasting power consumption, highlighting its potential utility in analyzing and predicting time-dependent data trends.

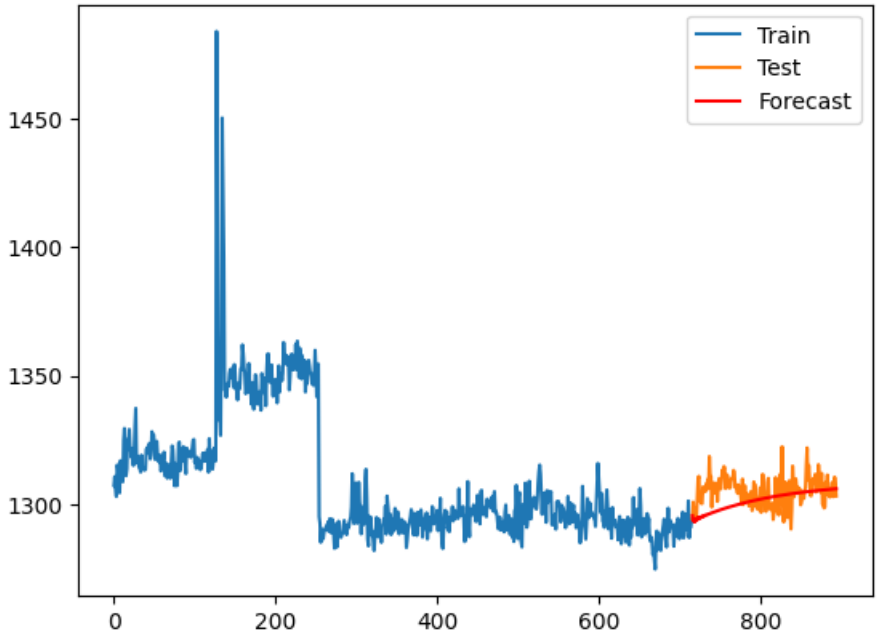


Figure 5.18 ARIMA Forecasting

5.4.2 Prophet

Prophet, a versatile time series forecasting library developed by Facebook, is employed in the provided code snippet to analyze and predict power consumption trends. This library offers an intuitive interface and powerful capabilities for forecasting time series data, making it a popular choice for various applications. The process begins with the initialization of the Prophet model, setting the stage for training on the dataset. Prophet operates on the principle of decomposing time series data into several components, including trend, seasonality, and holidays, thereby enabling accurate forecasting even with irregularly spaced data. Following initialization, the model is trained using the historical power consumption data. This involves learning the underlying patterns and trends in the data, which are crucial for making accurate predictions. Prophet employs an additive regression model that considers various factors contributing to the observed time series, such as trends and seasonal components. Once trained, the Prophet model is ready to generate forecasts for future time intervals. In this scenario, a future dataframe is created using the make_future_dataframe function, specifying the desired forecasting period. This dataframe serves as the basis for generating predictions beyond the existing dataset, allowing for insights into future power consumption trends. The predict method is then

applied to the trained model, utilizing the future dataframe to produce forecasts for each timestamp in the specified period. These forecasts encompass not only the expected power consumption values but also uncertainty intervals, providing valuable insights into the reliability of the predictions.

Finally, the forecasted data is visualized using the plot_plotly function, which generates an interactive plot showcasing both historical power consumption data and the forecasted values. This visualization aids in understanding the predicted trends and enables stakeholders to make informed decisions based on the projected power consumption patterns.

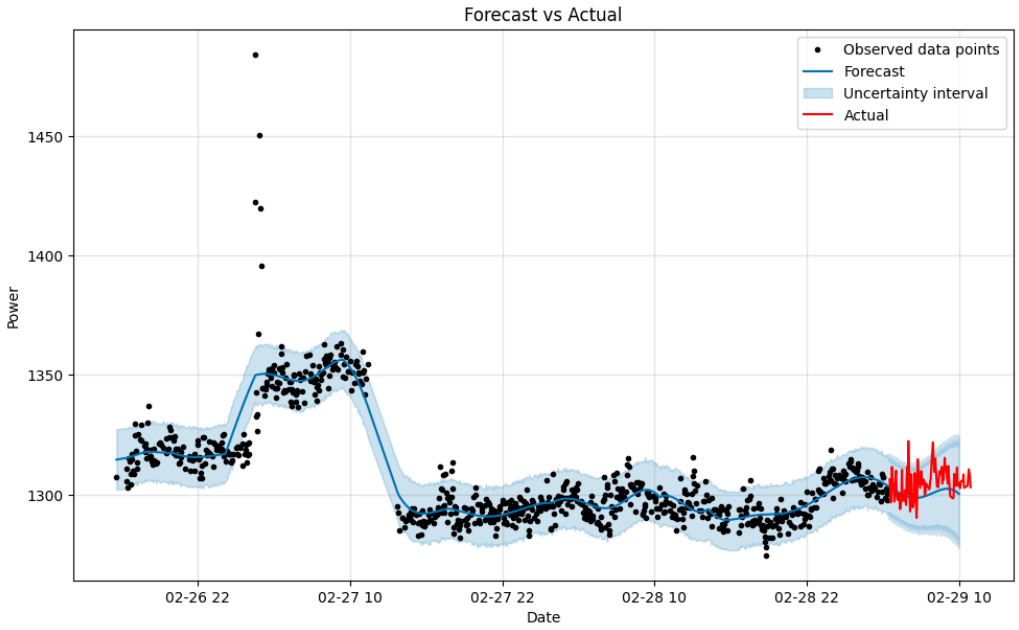


Figure 5.19 Prophet Forecasting

These forecasting techniques allowed us to anticipate the future trends and patterns in each parameter, enabling proactive maintenance interventions to prevent system failures and downtime. By leveraging historical data and time series analysis, we could identify potential anomalies and deviations from normal operating conditions, facilitating timely corrective actions and maintenance activities.

CHAPTER 6
RESULTS AND SNAPSHOTS

CHAPTER 6

RESULTS AND SNAPSHOTS

The data pipeline successfully integrates various stages, beginning with importing CSV files for preprocessing. This ensures data integrity by eliminating duplicates and handling null values. Leveraging pre-trained ensemble machine learning models enhances classification accuracy, enabling precise identification of pump system states. Subsequently, the classified data is seamlessly stored in a SQL database server, facilitating efficient management and retrieval. The visualization aspect, implemented through Power BI dashboards, provides stakeholders with intuitive insights for informed decision-making.

The dashboard encompasses two layouts: one for actual data and the other for forecasted values. The actual layout offers conditional monitoring of parameters such as power, vibration, and temperature, represented by color-coded status gauge meters. Green signifies normal operation, yellow indicates an alarm state, and red denotes shutdown.

In the forecasting layout, two lines illustrate the actual and predicted values, respectively. The solid line represents historical data, while the dotted line depicts future predictions generated by the Prophet model. Notably, the predictive capability of the model expands as historical data accumulates, allowing for more accurate long-term forecasts. The pipeline's effectiveness lies in its comprehensive approach, seamlessly transitioning from data preprocessing to visualization. By integrating machine learning models and database management, it empowers stakeholders with actionable insights to optimize pump system maintenance and operation. This holistic solution enhances decision-making processes by providing real-time monitoring, anomaly detection, and predictive analytics. Furthermore, the pipeline's adaptability enables scalability as data volumes grow, ensuring sustained performance and relevance over time. The insights derived from the pipeline serve as a foundation for proactive maintenance strategies, enabling organizations to minimize downtime, optimize resource allocation, and maximize operational efficiency. Overall, the results underscore the pipeline's significance in augmenting data analysis capabilities and driving informed decision support in pump system management.

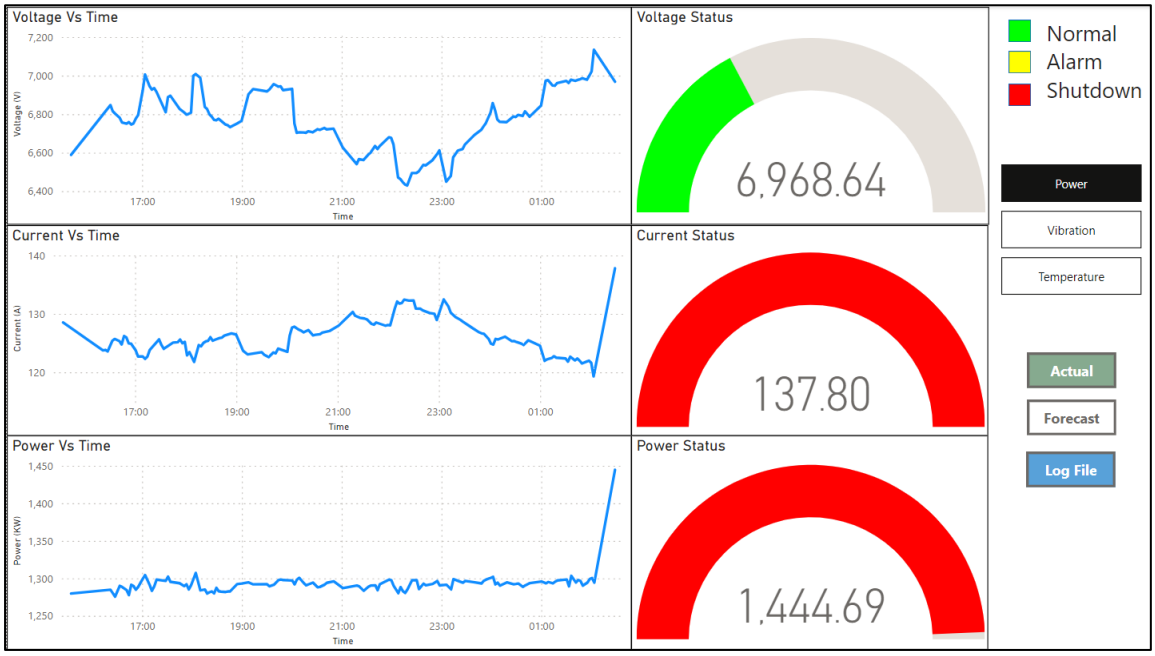


Figure 6.1 Dashboard showing actual status of Pump Power Consumption

Figure 6.2 Dashboard showing actual status of Pump Vibration

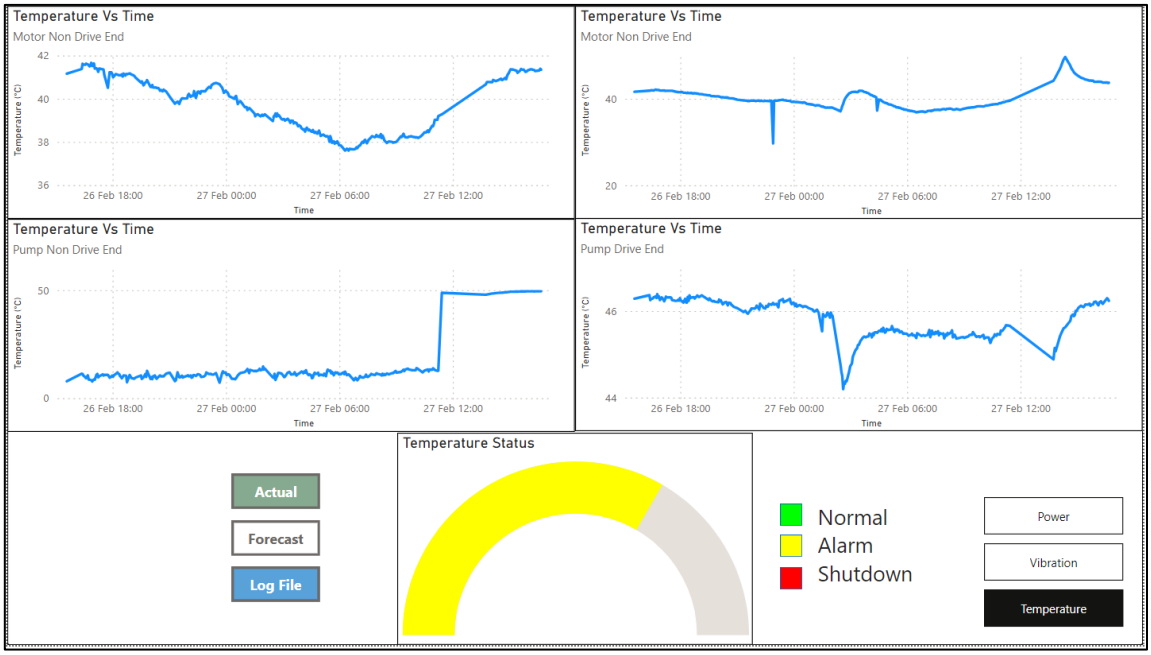


Figure 6.3 Dashboard showing actual status of Pump Temperature

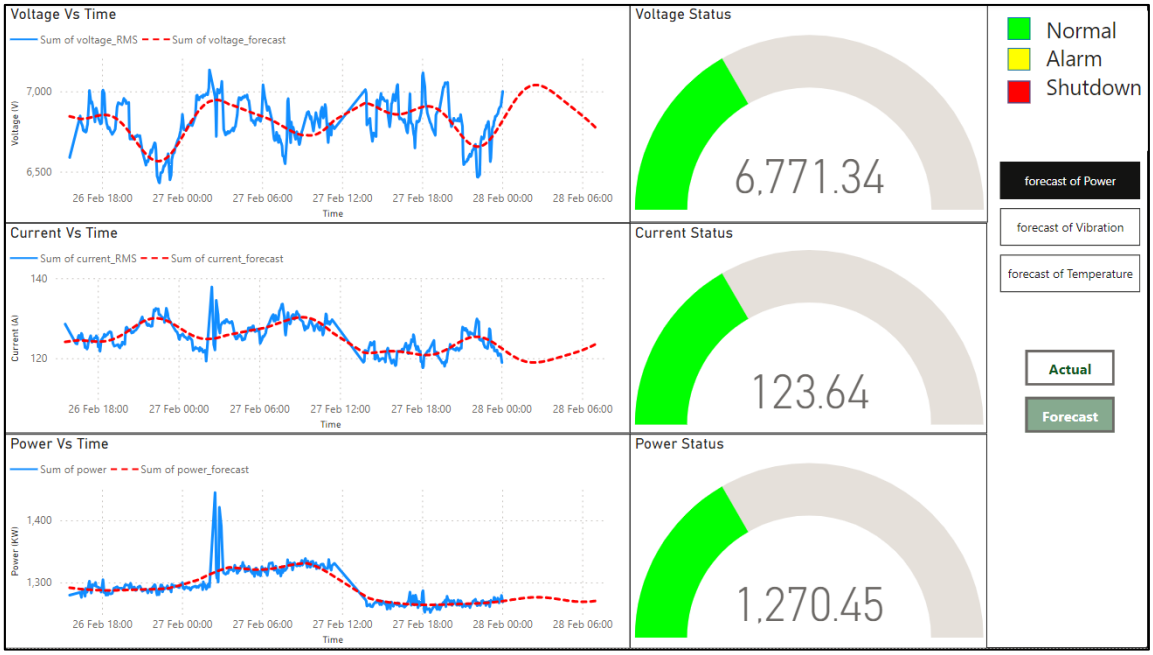


Figure 6.4 Dashboard showing forecasted status of Pump Power Consumption

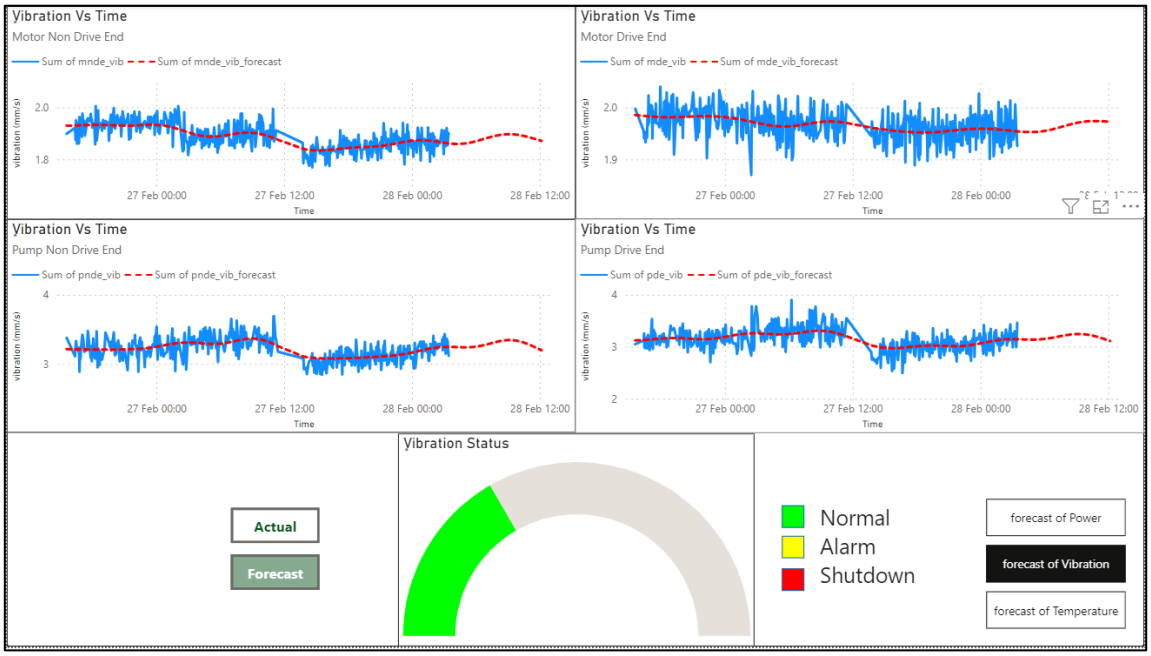


Figure 6.5 Dashboard showing forecasted status of Pump Vibration

Figure 6.6 Dashboard showing forecasted status of Pump Temperature

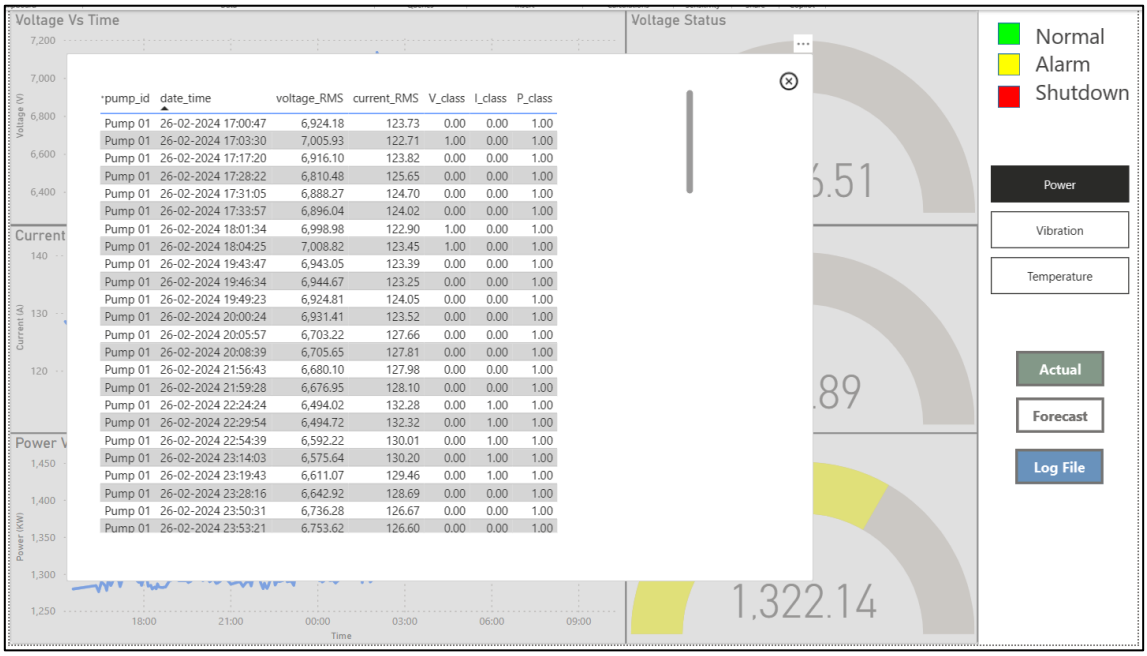


Figure 6.7 Dashboard showing log file of alarms

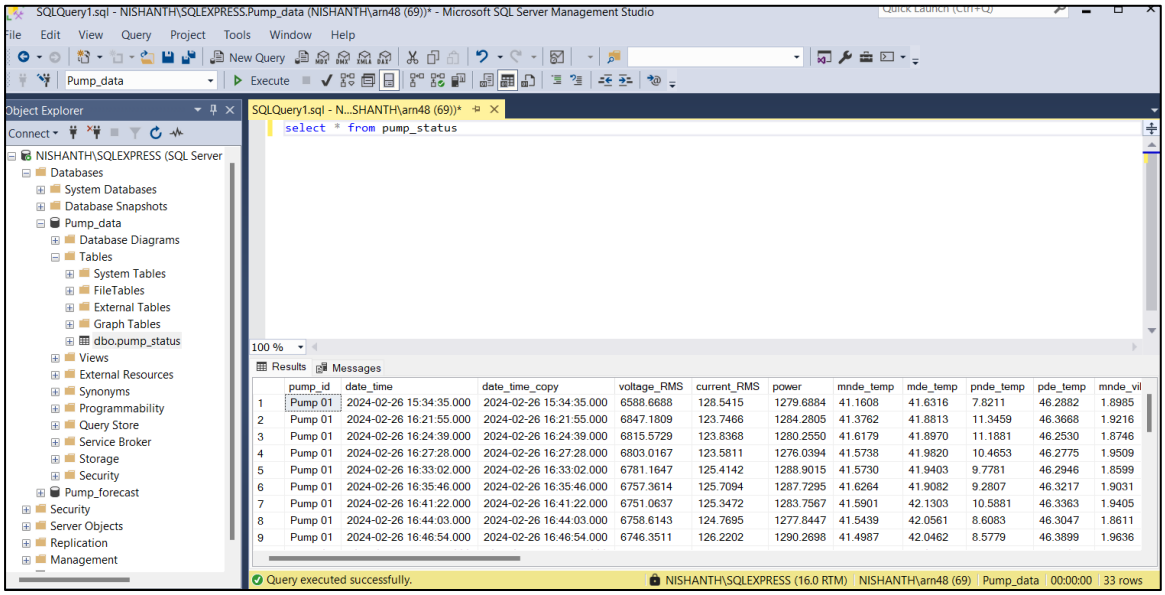


Figure 6.8 Glimpse of data uploaded to SQL database server

CONCLUSION

In conclusion, our comprehensive analysis and implementation of predictive maintenance techniques for pump systems have yielded promising results. Through exploratory data analysis (EDA), we gained valuable insights into the temporal patterns and relationships between key parameters, such as voltage, current, and power consumption. Leveraging K-means clustering, we successfully categorized pump states into normal, alarm, and shutdown categories, facilitating proactive maintenance interventions. By labeling the dataset based on predefined threshold values, we enabled the classification of pump conditions, further enhancing our predictive models' accuracy. Integration of diverse machine learning models and ensemble methods allowed us to effectively classify system states and mitigate individual model biases. Additionally, our forecasting models, including ARIMA, VAR, and Prophet, enabled us to anticipate future trends and deviations from normal operating conditions, empowering stakeholders with actionable insights for timely maintenance interventions. Overall, our approach holds significant potential for enhancing operational efficiency, minimizing downtime, and optimizing equipment uptime in industrial pump systems, contributing to improved reliability and performance.

REFERENCES

[1] Abdalla, Ramez & Samara, Hanin & Perozo, Nelson & Paz, Carlos & Jaeger, Philip. (2022). Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs). ACS Omega. 7. 10.1021/acsomega.1c05881.

[2] Amihai, Ido & Gitzel, Ralf & Kotriwala, Arzam & Pareschi, Diego & Subbiah, Subanatarajan & Sosale, Guru. (2018). An Industrial Case Study Using Vibration Data and Machine Learning to Predict Asset Health. 178-185. 10.1109/CBI.2018.00028.

[3] Dr. Sharda Chhabria & Rahul Ghata & Varun Mehta & Ayushi Ghosekar & Manasi Araspure & Nandita Pakhid. (2022). Predictive Maintenance Using Machine Learning on Water Pump.

[4] P. G. Kini, R. C. Bansal and R. S. Aithal, "Performance Analysis of Centrifugal Pumps Subjected to Voltage Variation and Unbalance," in IEEE Transactions on Industrial Electronics, vol. 55, no. 2, pp. 562-569, Feb. 2008, doi: 10.1109/TIE.2007.911947.

[5] M. Sanayha and P. Vateekul, "Fault detection for circulating water pump using time series forecasting and outlier detection," 2017 9th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 2017, pp. 193-198, doi: 10.1109/KST.2017.7886095.

[6] Jafar Zarei, Mohammad Amin Tajeddini, Hamid Reza Karimi,Vibration analysis for bearing fault detection and classification using an intelligent filter, Mechatronics, Volume 24, Issue 2, 2014, Pages 151-157, ISSN 0957-4158, doi.org/10.1016/j.mechatronics.2014.01.003.

[7] Sunilkumar. 2021. Predictive Maintenance of Pumps. Blog in Medium. https://medium.com/@suniaidvpr/predictive-maintenance-of-pumps-7c358f0efe68

DrillBit

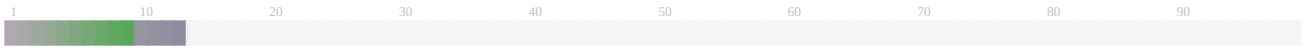
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

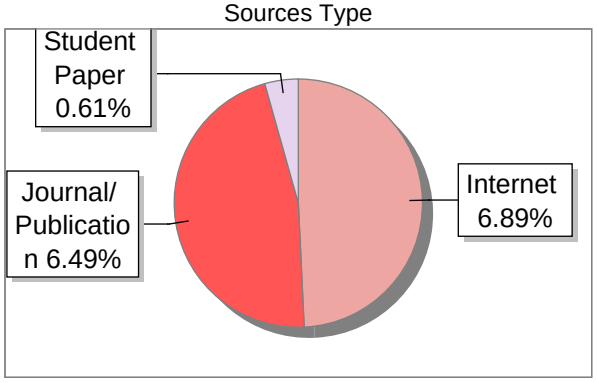
Author Name	Manthan Prasad
Title	Pump Anomaly Detection Using Machine Learning Techniques
Paper/Submission ID	1846795
Submitted by	libraryjssate@gmail.com
Submission Date	2024-05-21 21:48:20
Total Pages, Total Words	57, 9950
Document type	Project Work

Result Information

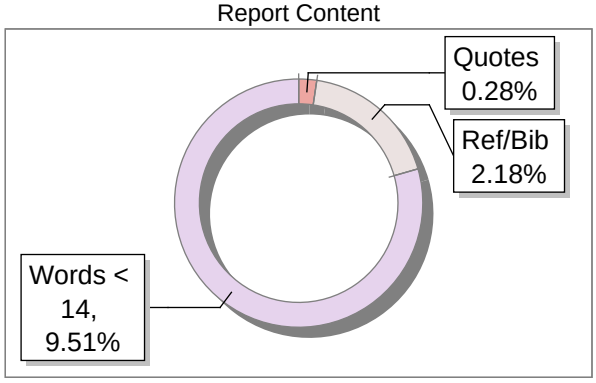
Similarity **14 %**



Sources Type



Report Content



Exclude Information

Quotes	Not Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

14	99	B
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.ijnrd.org	1	Publication
2	REPOSITORY - Submitted to Kalinga University, Raipur on 2024-03-19 18-01	<1	Student Paper
4	www.smec.ac.in	<1	Publication
6	arxiv.org	<1	Publication
7	businessdocbox.com	<1	Internet Data
8	ijrpr.com	<1	Publication
9	www.ncbi.nlm.nih.gov	<1	Internet Data
10	A Robust and Sparse Process Fault Detection Method Based on RSPCA by Peng-2019	<1	Publication
11	aben.springeropen.com	<1	Internet Data
12	datasciencedojo.com	<1	Internet Data
13	mdpi.com	<1	Internet Data
14	machinelearningmastery.com	<1	Internet Data
15	mdpi.com	<1	Internet Data

16	www.mdpi.com	<1	Internet Data
17	pecteam.in	<1	Publication
18	mdpi.com	<1	Internet Data
19	neurosys.com	<1	Internet Data
21	www.mdpi.com	<1	Internet Data
22	Enhanced stability, bistability, and exceptional points in saturable a by Zhiyenbayev-2019	<1	Publication
23	Formation, development, and propagation of a rare coastal coccolithop, by Matson, Paul G. Wa- 2019	<1	Publication
24	Thesis submitted to shodhganga - shodhganga.inflibnet.ac.in	<1	Publication
25	worldwidescience.org	<1	Internet Data
26	www.diva-portal.org	<1	Publication
27	www.mdpi.com	<1	Internet Data
28	www.ncbi.nlm.nih.gov	<1	Internet Data
29	www.simplilearn.com	<1	Internet Data
30	www.thefreelibrary.com	<1	Internet Data
31	www.ukcdr.org.uk	<1	Publication
32	aben.springeropen.com	<1	Internet Data
33	moam.info	<1	Internet Data
34	moam.info	<1	Internet Data

35	REPOSITORY - Submitted to Ayya Nadar Janaki Ammal College on 2024-03-27 15-38	<1	Student Paper
36	www.qmul.ac.uk	<1	Internet Data
37	ajod.org	<1	Publication
38	dspace.lib.cranfield.ac.uk	<1	Internet Data
39	jbe.tums.ac.ir	<1	Publication
40	moam.info	<1	Internet Data
41	onix-systems.com	<1	Internet Data
42	Workshop Synthesis Smart card data, new methods and applications for public tra by Chandesris-2018	<1	Publication
43	Editorial IEEE Transactions on Human Machine Systems Year in Review by Bass-216	<1	Publication
44	intl.finebi.com	<1	Internet Data
45	medium.com	<1	Internet Data
46	Thesis Submitted to Shodhganga Repository	<1	Publication
47	ugspace.ug.edu.gh	<1	Publication
48	www.mdpi.com	<1	Internet Data
49	www.ncbi.nlm.nih.gov	<1	Internet Data
50	citeseerx.ist.psu.edu	<1	Internet Data
51	moam.info	<1	Internet Data
52	nature.com	<1	Internet Data

53	www.sciencepublishinggroup.com	<1	Publication
54	IEEE 2015 24th International Conference on Computer Communication an by	<1	Publication
55	azdoc.site	<1	Internet Data
56	byjus.com	<1	Internet Data
57	clutejournals.com	<1	Publication
58	CoNeSec Track Report IEEE Conference Publicationby Grzegorz Koaczek 2017- ieeeexplore.org	<1	Publication
59	Developing grey prediction with Fourier series using genetic algorithms for tour by Hu-2020	<1	Publication
60	ejournal.undip.ac.id	<1	Internet Data
61	eprints.covenantuniversity.edu.ng	<1	Internet Data
63	journal2.um.ac.id	<1	Internet Data
64	moam.info	<1	Internet Data
65	pandas.pydata.org	<1	Publication
66	springeropen.com	<1	Internet Data
67	theses.hal.science	<1	Publication
68	Thesis Submitted to Shodhganga Repository	<1	Publication
69	visionpdf.com	<1	Internet Data
70	www.dx.doi.org	<1	Publication
71	www.frontiersin.org	<1	Publication

72	www.mdpi.com	<1	Internet Data
73	armypubs.army.mil	<1	Publication
74	arxiv.org	<1	Publication
75	A comprehensive evaluation of multicategory classification methods for fault cla by V-2009	<1	Publication
76	Boa Ultra-Large-Scale Software Repository and Source-Code Mining by Dyer-2015	<1	Publication
77	Comparison of Multilayer Perceptron and Long Short-Term Memory for Plant Paramet by Bae-2019	<1	Publication
78	docplayer.net	<1	Internet Data
79	docplayer.net	<1	Internet Data
80	ebin.pub	<1	Internet Data
81	Estimating Polling Accuracy in Multiparty Elections Using Surveybias by Arzheimer-2016	<1	Publication
82	eurekaselect.com	<1	Internet Data
83	Evaluation of fracability and screening of perforation interval for tight sandst by Guo-2015	<1	Publication
84	journal.umy.ac.id	<1	Publication
85	journal2.um.ac.id	<1	Internet Data
86	mdpi.com	<1	Internet Data
88	moam.info	<1	Internet Data
89	Quantifying Operational Disruptions as Measured by Transportation Network Reliab by Pennetti-2020	<1	Publication

90	Quantifying Operational Disruptions as Measured by Transportation Network Reliab by Pennetti-2020	<1	Publication
91	Recent Advances in the Internet of Medical Things (IoMT) Systems Security by Ghubaish-2020	<1	Publication
92	repository.up.ac.za	<1	Publication
93	springeropen.com	<1	Internet Data
94	thedesignengineering.com	<1	Internet Data
95	Thesis submitted to dspace.mit.edu	<1	Publication
96	Thesis Submitted to Shodhganga Repository	<1	Publication
97	The effect of length of follow-up on substantial clinical benefit thr, by Spurgas, Morgan P.- 2019	<1	Publication
98	Trends in non-stationary signal processing techniques applied to vibration analy by Um-2017	<1	Publication
99	www.dx.doi.org	<1	Publication
100	www.dx.doi.org	<1	Publication
101	www.dx.doi.org	<1	Publication
102	www.dx.doi.org	<1	Publication
103	www.naun.org	<1	Publication
104	www.priorilegal.com	<1	Internet Data