PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA MINISTRY HIGHER EDUCATION AND SCIENTIFIC RESEARCH KASDI MERBAH UNIVERSITY OUARGLA

Faculty of New Technologies of Information and Communication Department of Electronics and Telecommunication

Final Dissertation

Thesis for Master Degree in

Automatic and Systems

Thesis

Development of a Machine Learning-Based Predictive Maintenance System for Maintenance Enhancement

---

Submitted by:

DOUH AIMEN SALIM

Defense Committee:

Dr. Benarabi Bilal	President	UKM Ouargla
Dr. Rouabah Boubakeur	Examiner	UKM Ouargla
Dr. Achbi Mohammed Said	Supervisor	UKM Ouargla
Mm. Ababsa Amani	Co- Supervisor	UKM Ouargla

Academic year: 2024/2025

Dedication

To the most important people in my life, I dedicate this thesis to..

my mother,
Your sincere prayers, kind heart, and constant support have been the light that guided me through every step of this journey.

my father,
Your strength, quiet support, and unconditional love gave me the courage to keep going, even in the hardest times.

To both of you
This thesis is dedicated to you, with love, gratitude, and deep respect that words can never fully express.

my dear siblings,
You have been my strength and motivation through every stage of this path. Your support means the world to me.

And to my closest friends,
Thank you for walking beside me with your honest hearts, true friendship, and constant encouragement.
This work carries your presence in every page.

Acknowledgments

I would like to express my sincere thanks and appreciation to my supervisor **, Achbi Mohammed Said,** for his valuable guidance, support, and patience throughout this project. His advice played a major role in completing this work.

I am also deeply grateful to **Ababsa Amani**, for her help, encouragement, and continuous presence during every step of the process.

I would also like to thank **KASDI MERBAH UNIVERSITY OUARGLA** for providing a supportive learning environment and the academic resources that made this work possible.

My appreciation goes as well to all the professors who taught me and supported me throughout my academic journey.

Finally, to my friends and classmates thank you for your kind words, support, and presence.
You were an important part of this achievement, and I will always be thankful for that.

Abstract

This research utilized machine learning to build a predictive model for the GE MARK VIe industrial control system. The model aimed to estimate the G2.TTXD output based on three input variables: Inlet Pressure, Inlet Temperature, and IGV Position. Data gathered from the Hassi R'mel gas field was preprocessed and used to train and evaluate three regression algorithms: Linear Regression, Random Forest, and Support Vector Regression (SVR).

Model performance was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), along with analyses of residuals and value evolution. Among the models, SVR delivered the most accurate and reliable results.

These findings demonstrate the potential of machine learning to enhance industrial control systems through predictive maintenance, real-time monitoring, and increased operational efficiency key objectives of Industry 4.0.

Keywords: PdM, ML, Industry 4.0, GE Mark VIe, Speedtronic , SCADA / DCS, ToolboxST, ControlST, CIMPLICITY, OPC/ Modbus, Python, scikit-learn, SVR, Random Forest, Linear Regression, MSE, MAE, R^2, G2.TTXD, Inlet Pressure, Inlet, Temp IGV Position

Résumé

Cette recherche a utilisé l'apprentissage automatique (machine learning) pour construire un modèle prédictif destiné au système de contrôle industriel GE MARK VIe. Le modèle visait à estimer la sortie G2.TTXD à partir de trois variables d'entrée : la pression d'entrée, la température d'entrée et la position des IGV (Inlet Guide Vane). Les données recueillies sur le champ gazier de Hassi R'mel ont été prétraitées puis utilisées pour entraîner et évaluer trois algorithmes de régression : la régression linéaire, la forêt aléatoire (Random Forest) et la régression par vecteurs de support (SVR).

La performance des modèles a été évaluée à l'aide de l'erreur quadratique moyenne (MSE), de l'erreur absolue moyenne (MAE) et du coefficient de détermination (R^2), complétées par des analyses des résidus et de l'évolution des valeurs. Parmi les modèles testés, le SVR a fourni les résultats les plus précis et fiables.

Ces résultats démontrent le potentiel de l'apprentissage automatique pour améliorer les systèmes de contrôle industriels à travers la maintenance prédictive, la surveillance en temps réel et l'augmentation de l'efficacité opérationnelle des objectifs clés de l'Industrie 4.0.

Mots-clés : PdM, ML, Industrie 4.0, GE Mark VIe, Speedtronic, SCADA / DCS, ToolboxST, ControlST, CIMPLICITY, OPC / Modbus, Python, scikit-learn, SVR, Forêt Aléatoire, Régression Linéaire, MSE, MAE, R^2, G2.TTXD, Pression d’entrée, Température d’entrée, Position IGV.

ملخص

استخدم هذا البحث تقنيات التعلم الآلي لبناء نموذج تنبؤي لنظام التحكم الصناعي GE MARK VIe يهدف النموذج إلى تقدير قيمة خرج G2.TTXD بناءً على ثلاثة متغيرات إدخال هي: ضغط الدخول، درجة حرارة الدخول، وموضع موجهات الدخول (IGV) .

تمت معالجة البيانات التي جُمعت من حقل الغاز بحاسي الرمل واستخدامها لتدريب وتقييم ثلاث خوارزميات انحدار: الانحدار الخطي، الغابات العشوائية، والانحدار بدعم المتجهات .(SVR) تم تقييم أداء النماذج باستخدام مقاييس متوسط الخطأ التربيعي(MSE) ، ومتوسط الخطأ المطلق(MAE) ، ومعامل التحديد(R^2) ، إلى جانب تحليل البواقي وتطور القيم .

من بين هذه النماذج، قدم نموذج SVR النتائج الأكثر دقة وموثوقية .تُظهر هذه النتائج إمكانيات التعلم الآلي في تعزيز أنظمة التحكم الصناعية من خلال الصيانة التنبؤية، المراقبة في الوقت الحقيقي، وزيادة الكفاءة التشغيلية، وهي أهداف رئيسية لمفهوم الصناعة 4.0.

Contents

List of Figures

List of tables

General Introduction

In the rapidly evolving landscape of industrial automation, predictive maintenance (PdM) has emerged as a pivotal strategy for enhancing operational reliability, reducing unscheduled downtimes, and optimizing asset lifecycles. Among the myriad of industrial assets, gas turbines represent a particularly critical and complex class of machinery whose failures can result in substantial economic losses and safety hazards. The General Electric (GE) Mark VIe control system is widely deployed in managing the operation of gas turbines, providing real-time monitoring, control, and diagnostics. Despite its advanced functionalities, the system's current diagnostic capabilities largely rely on rule-based or threshold-triggered alerts, which may fail to detect subtle degradation patterns or predict imminent failures with high accuracy.

Recent advancements in machine learning (ML) have shown considerable promise in augmenting PdM strategies, offering the ability to model complex, nonlinear relationships within multivariate sensor data to detect anomalies, estimate remaining useful life (RUL), and forecast component failures. However, the integration of ML-based PdM systems into highly specialized industrial control platforms such as the GE Mark VIe remains relatively underexplored in the academic literature. Existing studies often focus on generic ML applications or simulations, with limited emphasis on the practical challenges and opportunities of deployment within proprietary industrial ecosystems. Additionally, many prior works lack access to real-world data from operational turbines, which is crucial for developing robust, generalizable models.

This study seeks to bridge a critical gap in predictive maintenance (PdM) by designing and evaluating a machine learning (ML)-driven solution specifically tailored for gas turbines managed by GE Mark VIe control systems. The primary objectives are to: (1) develop a robust ML model capable of detecting early mechanical or operational anomalies using high-resolution sensor data from the Mark VIe platform; (2) evaluate the model's integration feasibility within the constraints of the existing industrial control architecture; and (3) validate its effectiveness through comprehensive testing on real-world historical data from operational gas turbines. A key contribution of this research lies in moving beyond generic PdM approaches toward a domain-specific, data-centric methodology that harnesses the predictive power of ML.

Emphasizing the importance of model evaluation, the study provides a detailed performance analysis that highlights the effectiveness, scalability, and reliability of the machine learning approach in predictive maintenance. The evaluation focuses on key metrics such as accuracy, responsiveness, and robustness across varying operational conditions, demonstrating the model's capacity to adapt to complex industrial environments. This thorough assessment underscores the feasibility of integrating ML into existing automation systems and its potential to enhance maintenance planning, resource optimization, and system availability. By delivering a validated and practical ML-based PdM solution, this research contributes to the advancement of intelligent maintenance strategies and supports the growing convergence of machine learning and industrial control systems—an increasingly vital area in the evolution of automation engineering.

CHAPTER 1: Description of the Hassi R'mel Field

.1 Introduction:

Algeria is one of the main natural gas-producing countries, possessing about 10% of the world's reserves, ranking fifth internationally. More than 50% of its known reserves are concentrated in the Hassi R'Mel field.

.2 Geographical Location of Hassi R'mel:

Hassi R'mel is located 522 km south of Algiers, between the wilayas of Laghouat and Ghardaïa, at an altitude of 760 meters. The landscape consists of a vast rocky plateau.
The Hassi R'mel field is situated on the Saharan platform near the northeastern edge of the Sahara. It is one of the largest gas fields in the world, with an elliptical shape covering an area of 3,500 km² (70×50). (See Figure I.1) [1].

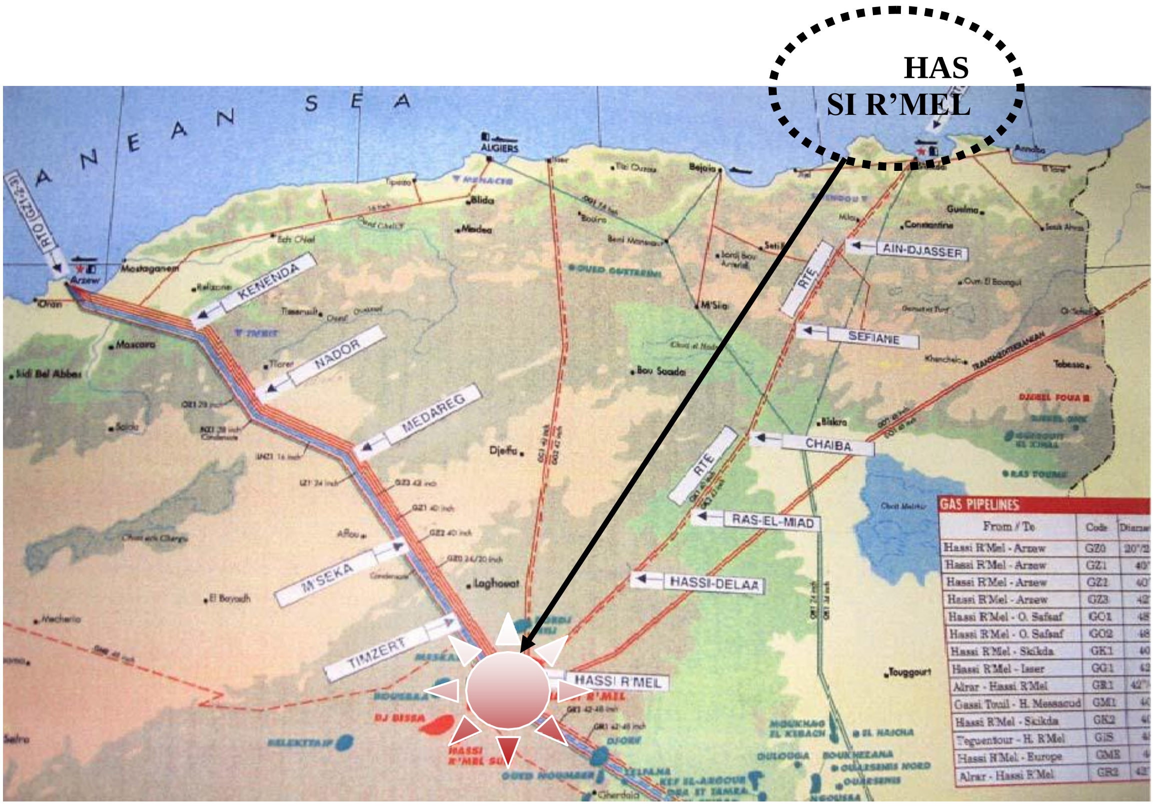

Figure 1.1: Geographical location of Hassi R'mel [1]

.3 Gas Installations of Hassi R'mel:

The overall layout of the gas installations located in the Hassi R'mel field is designed to enable the efficient exploitation of the reservoir and to maximize the recovery of liquids (condensate). Hassi R'mel is divided into three sectors [1]:

- **Northern Sector**
  Composed of a gas processing module (MPP3) and a compression station.
- **Central Sector**
  Composed of three gas processing modules (MPP0, MPP1, and MPP4), the fluid storage and transfer center (for condensate and LPG), an Oil Treatment Center (CTH), and the common unit - phase B.

- **Southern Sector**
  Composed of a gas processing module (MPP2), the southern compression station, Djebel Bissa, and HR South. The four gas treatment units 1, 2, 3, and 4 each have a capacity of 60 m³/day. They were built in 1970/1980. However, MPP0 has a capacity of 30 million m³/day and Djebel Bissa has a capacity of 6 m³/day.

CTH2, CTH4, CTH1, CTH3: Gas treatment units
BOOSTING Station: Oil treatment center
Gas Re-injection Station:
Oil Loop
Southern Zone – Central Zone – Northern Zone: Associated Gas Recovery Station [1].

Another compression station, called Boosting, is currently in operation. It is designed to increase the gas inlet pressure for the processing modules to ensure the continuous exploitation of the gas field. (See Figure I.2) [1].

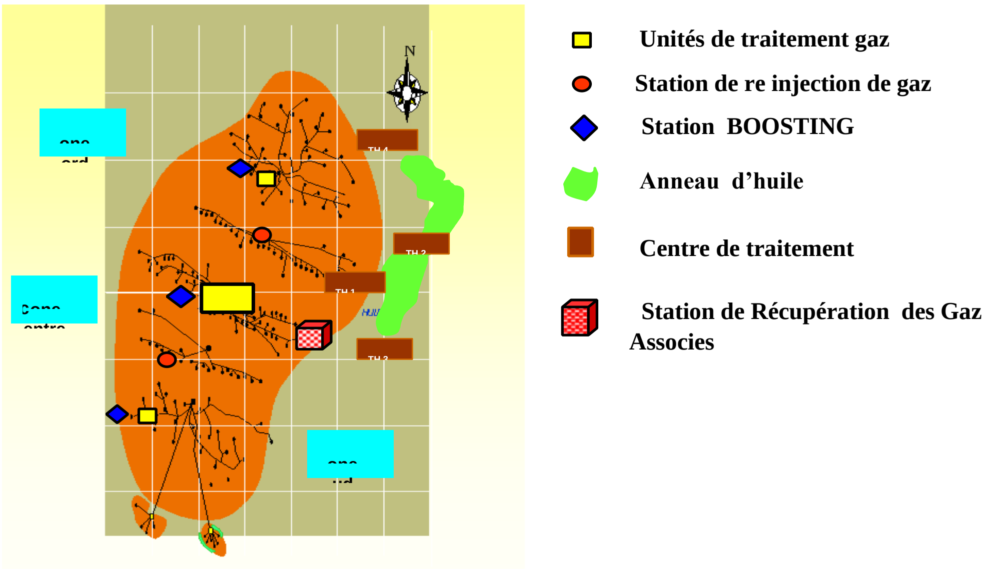

Figure 1.2: Gas installations of Hassi R'mel [1]

.4 Description of the North Compression Station:

.4.1 General Information:

The north and south compression stations are designed to compress the gas coming from gas treatment facilities (MPP0, MPP1, MPP2, MPP3, and MPP4) and reinject it into the underground reservoir (the gas field of Hassi R'mel) [2].
The purpose of reinjection is to:

- Maintain the gas pressure in the national and international distribution network at 70 bars.
- Avoid flaring the gas and losing it unnecessarily.
- Keep the reservoir pressure high to recover the maximum amount of liquids (condensate and LPG).

.4.2 Operating Characteristics of the Station:

- Gas pressure at the station inlet: 70 bars
- Gas temperature at the station inlet: 45°C
- Gas pressure at the station outlet: 350 bars
- Gas temperature at the station outlet: 100°C

.4.3 Description of Station Operation:

The compression station is designed to compress gas up to 350 bars.
Compression is carried out in two stages with intermediate cooling. The low pressure stage compresses the gas up to 150 bars, and the high pressure stage compresses it up to 350 bars.
The station includes 9 parallel compression lines (see Figure I.3).
Each stage uses nine "Barrel" type centrifugal compressors installed in parallel.
An intermediate manifold is provided, to which the discharge outlets of the nine first-stage compressors and the suction inlets of the nine second-stage compressors are connected, as well as a final manifold that receives the discharge of the nine second-stage compressors (see Figure I.3).
Each compressor is driven by a dual-shaft gas turbine, model MS 5002B (manufactured by NUOVO PIGNONE under GENERAL ELECTRIC license), through a speed increaser.
The compression units are housed in acoustic enclosures equipped with a special ventilation system to reduce the heat generated by the machines. These units are grouped into four prefabricated buildings ventilated by fans placed on the roof [2].

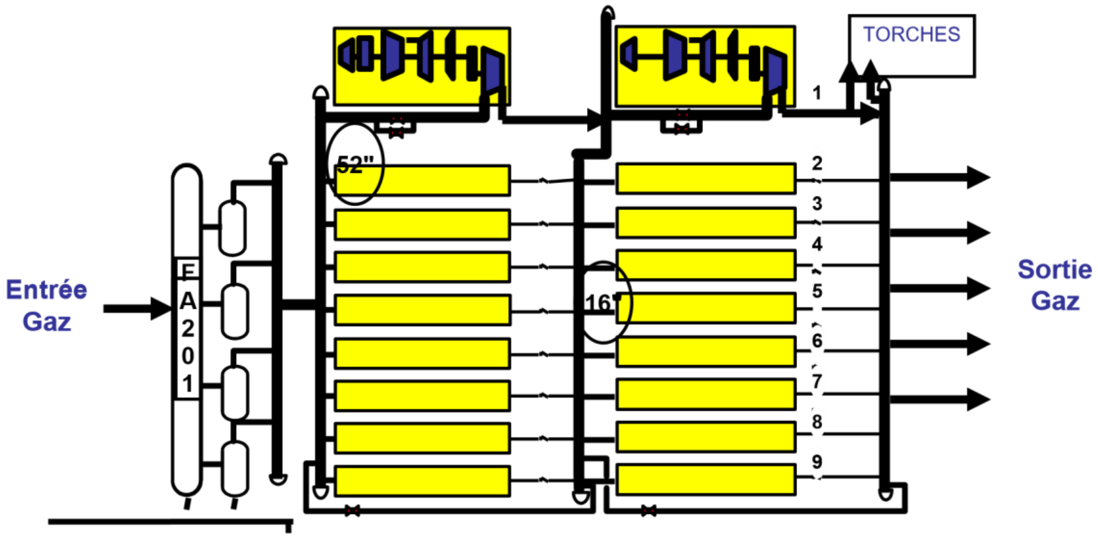

Figure 1.3: Northern and Southern Compression Station: (NCS and SCS).

.4.4 Compression Unit:

.4.4.1 First Stage Turbo Compressor :

- Turbine: Model MS 5002B, built by NUOVO PIGNONE under license from GENERAL ELECTRIC, simple cycle with two shafts.
- Gearbox: Type GB-47 with parallel axes, manufactured by MAAG.

- Compressors: Model BCL 405/a, built by NUOVO PIGNONE.

.4.4.2 Second Stage Turbo Compressor :

Turbine: Identical to that of the first stage.

Gearbox: Identical to that of the first stage.

Compressor: Model BCL 405/c, built by NUOVO PIGNONE, BARREL type for very high pressure.

.5 Process Description

.5.1 Electrical Power Supply

The station is powered by two 30 KW lines from SONELGAZ, each arriving at the electrical substation (SSE), where the voltage is stepped down to 5.5 KV using two 8 MVA transformers.
The substation supplies five transformer cabins known as power centers, equipped with 1600 KW and 1000 KVA transformers, which step the voltage down to 380V to power the various receivers.
In case of a power outage on the SONELGAZ network, the station is equipped with an electric turbogenerator (TGE) with a capacity of 3440 KVA, which can supply the station's priority receivers and thus prevent shutdown [2].

.5.2 Description of the Gas Circuit Process

A - First Stage Suction Section and Scrubbers

The gas enters the station at a pressure of 70 bars through two 36″ connections, which connect to the 52″ suction manifold. On this line is installed the V201 isolation ball valve with a pneumatic actuator that uses line gas as its driving gas (for safety reasons).
To pressurize the line downstream of V201 and allow its opening, the automatic pressurization valve V202 is installed, which introduces gas into the line until the pressure switch PDSL 1-20 indicates that the pressure difference between the upstream and downstream of V201 is less than 1.5 bars, thereby allowing the V201 to open.

Next, the gas enters the empty vessel FA-201, which acts as a pre-separator to remove any large quantities of liquid carried by the gas flow. This gas then flows into parallel-installed baffle scrubbers capable of separating small liquid particles carried by the gas. These scrubbers are FA 202-203-204-205.
At the scrubber outlet, the gas flow is measured by the flow meter FA-201, after which it is sent to the machines. All separated liquid discharges are collected in the 6″ hydrocarbon manifold and sent to the end of the station for possible recovery [2].

B - First Stage Compression Unit

The gas enters the centrifugal compressor where it is compressed to a pressure of 150 bars. The compressor is driven by a gas turbine, type MS 5002B, through a two-horizontal-shaft gearbox [2].

C - First-Stage Compression Cooler

The gas compressed by the first-stage compressor is cooled in the interstage cooler (e.g., EA 221), where the cooling element is ambient air [2].

D - First-Stage Separator

Downstream of the first stage cooler is the first stage discharge separator, of the baffle type, used to separate any water vapor, hydrocarbons, or condensate [2].

E - Interstage Manifold

All the discharges from the 9 first stage compressors are connected to the 16" interstage manifold installed on the rack next to the 52" suction manifold [2].

F - Second-Stage Compression Unit

The gas enters the second stage compressor, where it is compressed to a pressure of 350 bars; the compressor is driven by a gas turbine identical to that of the first stage, via a two-shaft horizontal gearbox [2].

G - Second-Stage Gas Cooler

The gas compressed by the second stage compressor is cooled in the final cooler (e.g., EA 222), where the cooling element is ambient air [2].

H - Final Manifold

The discharge piping of the second stage compressors and the gas supply piping to the wells are connected to the station's final manifold (20" nominal diameter) [2].

.6 Conclusion:

This chapter presented a broad overview of the Hassi R'mel field and the process of the North Compression Station. Nonetheless, our main focus is on the turbocompressor, specifically the MS 5002B gas turbine and the BCL 405 A and C centrifugal compressors. The next chapter will detail how the turbocompressor functions.

CHAPTER 2: Overview of Gas Turbine

.1 Definition

A gas turbine is a rotary motion, internal combustion engine. Equipped with an air compressor and combustion chambers, it is capable of producing a high temperature, high pressure fluid that expands through the turbine stages to deliver mechanical energy for driving a receiving machine.

Their significant role in energy conversion processes is supported by the wide range of their applications across various fields. The preference for gas turbines in diverse uses is primarily based on their operational flexibility, reliability, and ease of maintenance compared to their traditional competitors steam turbines and reciprocating engines [3].

From a constructional point of view, there are two types of gas turbines:

A-Single-shaft turbine: The compressor and turbine sections are mounted on the same shaft, allowing them to rotate at the same speed. This type is used in applications that do not require variable speeds, such as driving generators for electricity production.

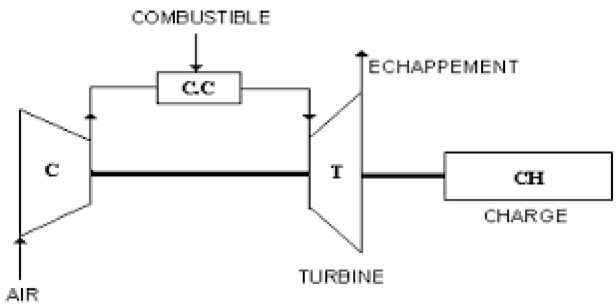

Figure 2.4: Diagram of the single-shaft turbine

B-Twin-shaft turbine: Unlike the single shaft gas turbine, the two turbine sections are not mechanically connected, allowing them to operate at different speeds. This type is used in applications requiring a wide range of speed variation, such as compressor drive systems.

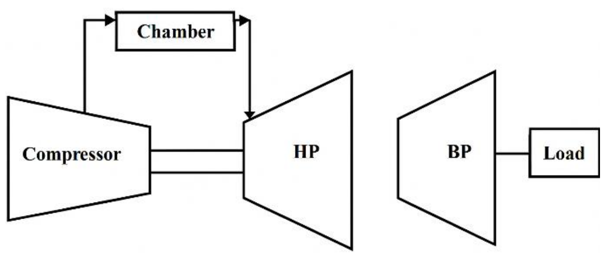

Figure 2.5: Diagram of the twin-shaft turbine

.2 MS 5002B Gas Turbine:

.2.1 Overview of the MS 5002B Gas Turbine:

The MS 5002B NUOVO-PIGNONE gas turbine is a two-shaft drive turbine operating on a simple cycle. It is used to drive a centrifugal compressor.
The portion of a gas turbine used for mechanical drive is the part where fuel and air are utilized to generate power on the shaft.
This turbine has two mechanically independent turbine wheels.
The high-pressure first-stage turbine wheel drives the air compressor rotor, which is a sixteen-stage axial type, as well as the shaft powering the auxiliary systems (lubrication and hydraulic pumps).
The second-stage turbine wheel, or low-pressure stage, drives the load (centrifugal compressor).
The two turbine wheels are not connected, allowing them to rotate at different speeds to adapt to varying load conditions.
This turbine model installed at Hassi R'mel operates on natural gas [3].

.2.2 Operating Principle:

A gas turbine operates as follows:

- It draws in air from the surrounding environment.
- It compresses the air to a higher pressure.
- It increases the energy level of the compressed air by injecting and burning fuel in a combustion chamber.
- It directs the high-pressure and high-temperature air to the turbine section, which converts thermal energy into mechanical energy to rotate the shaft. This serves both to:
    - Provide useful energy to the driven machine, coupled via a coupling device, and
    - Supply the necessary energy for air compression, which occurs in a compressor directly connected to the turbine section.
- It releases the low-pressure gases and resulting heat to the atmosphere following the energy transformation. (See Figure II.5)

.2.3 The Main Sections of the MS 5002B Gas Turbine:

A gas turbine is divided into five main sections:

- **Inlet**
- **Axial Compressor**
- **Combustion Chambers**
- **Turbine Section**
- **Exhaust**

.2.4 Inlet:

Gas turbines consume a large amount of air for combustion and cooling of internal components. This air must be filtered to prevent the penetration of particles that could, over time, erode the equipment. The turbine’s inlet is a chamber or compartment that houses the filters and is connected to the turbine's intake casing. This system combines air filtration and noise reduction functions [3].

.2.5 Axial Compressor:

The compressor is of the axial flow type. The axial flow system produces high air flow rates, which are necessary to obtain high useful power with compact dimensions. A compressor consists of a series of rotating blade stages (rotor) that increase the air velocity in terms of kinetic energy, alternated with fixed blade stages (stator), which convert the kinetic energy into higher pressure.

The number of compression stages is related to the structure of the gas turbine and, especially, to the compression ratio to be achieved. In the MS 5002B, there are 16 stages. On the intake side of the compressor, there are variable inlet vanes (IGV), whose primary purpose is to direct the air supplied by the intake system toward the first stage of the rotor blades. Another important function of the IGVs is to ensure correct fluid-dynamic behavior of the compressor under different transient operating conditions (for example, during startup and shutdown), when, due to different operating speeds compared to the normal operating speed, the IGV opening angle can be adjusted; this serves to regulate the discharge pressure of the axial compressor.

The compressor also serves to provide a source of air necessary for cooling the walls of the guide vanes, blades, and turbine disks, which is achieved through internal channels in the gas turbine and external connecting piping [3].

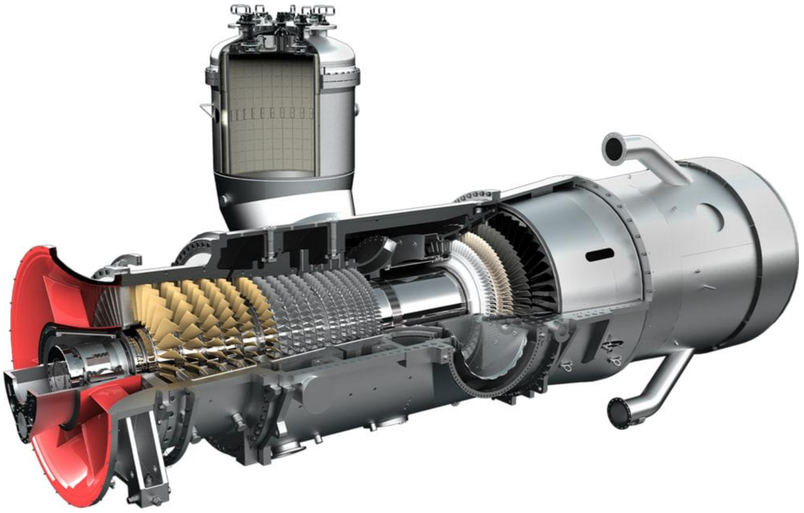

Figure 2.6: Axial Compressor

Torque and Work in Axial Compressors:

According to Euler's Turbomachinery Equation, the torque applied to a steady-flow axial compressor is directly related to the change in angular momentum of the fluid

$$\tau_a = \dot{m}\,(r_e\, v_e - r_i\, v_i)$$

Where:
τ_a: Torque applied
$\dot{m}$: Mass flow rate
r_e, r_i: Exit and inlet radii
v_e, v_i: Tangential velocity at exit and inlet respectively

The work done by the shaft (input power) is given by:

$$W_c = \omega\, \tau_a = \dot{m}\, \omega\, (r_e\, v_e - r_i\, v_i)$$

Where:
W_c: Compressor work

.2.6 Combustion Section:

After being compressed in the compressor, the air exits and enters the combustion chamber to partially participate in the combustion process, which provides a very high thermal energy input. The combustion system of the MS 5002B consists of twelve cylindrically shaped combustion chambers arranged symmetrically around a circumference. Each chamber contains a flame tube (combustion liner), a cap, a gas injector, and a transition piece.

In addition to these components, twelve interconnection tubes allow the flame to propagate between the twelve combustion chambers. Two spark plugs, located in flame tubes 1 and 12, initiate combustion, and two flame detectors, placed in flame tubes 3 and 10, ensure that combustion is occurring in all chambers. The hot gases resulting from combustion are directed to the first turbine wheel of the first stage via the transition pieces, which convert the cylindrical shape of the gas path into an annular shape suitable for the turbine wheel [3].

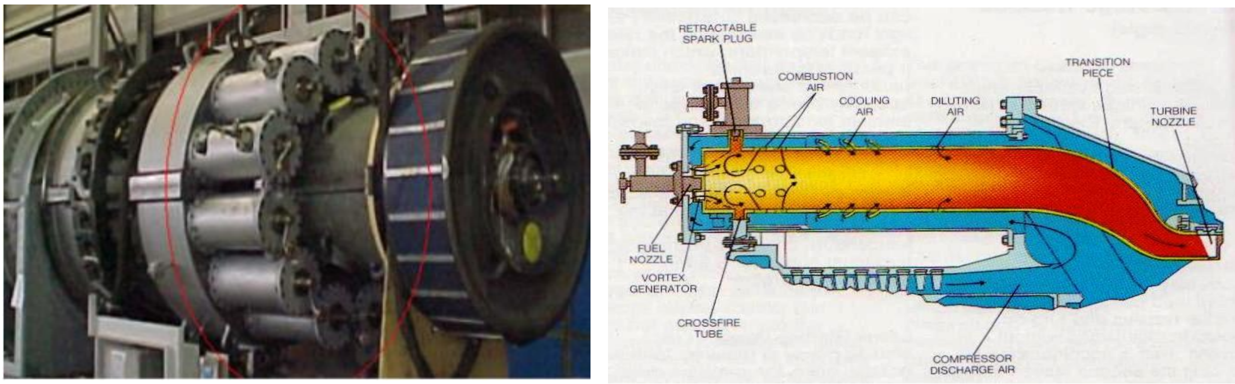

Figure 2.7: Combustion chambers

.2.7 Turbine Section:

After combustion, the hot compressed gases move toward the path of least resistance (the exhaust), passing through the HP (high-pressure) wheel to release a large portion of their stored energy. The trajectory of the gases in the HP wheel is tangential to the intrados profile of the blades to avoid shocks (which cause energy loss) and to achieve a maximum resulting torque.

The HP wheel is directly connected to the rotor of the axial compressor; this assembly is often referred to as the rotor. The expansion of gases in the HP wheel serves to rotate the axial compressor.

The gases exiting the HP wheel pass through the second-stage stator, which helps regulate the speed of the HP wheel and the exhaust temperature using its variable vanes. Then, a second expansion occurs in the LP (low-pressure) wheel, located just after the stator. The blades of the LP wheel are longer than those of the HP wheel in order to maximize the contact surface (for greater torque). They are supported at both ends to prevent bending.
The gases leaving the LP wheel are discharged into the atmosphere, and the resulting torque is used to drive the load (centrifugal compressor) [3].

Figure 2.8: The LP wheel

Figure 2.9: Axial compressor section and The HP wheel

.2.8 Exhaust:

The exhaust system serves to discharge into the atmosphere the gases resulting from the expansion in the turbine wheels [3].

CHAPTER 3: Overview of GE Mark VIe

.1 Introduction

The rapid deployment of industrial automation and control systems has largely made complex processes more reliable, safe, and efficient, the energy field being the best example. One of these advanced systems is the Mark VIe**,** developed by General Electric. Mark VIe is a high performance, flexible control system that is appropriate for controlling turbines of all sizes. With its modular architecture, powerful features, and extensive interfacing functionality, it is one of the most prolific control systems used in both power generation and many other industrial settings.[4]

This chapter explains the Mark VIe system in detail, including its structure, the hardware and software it uses, how it communicates, what it can do, and how it is used in industrial automation. Knowing how this system works is important for engineers and researchers who work on predictive maintenance, system reliability, and industrial control.[4]

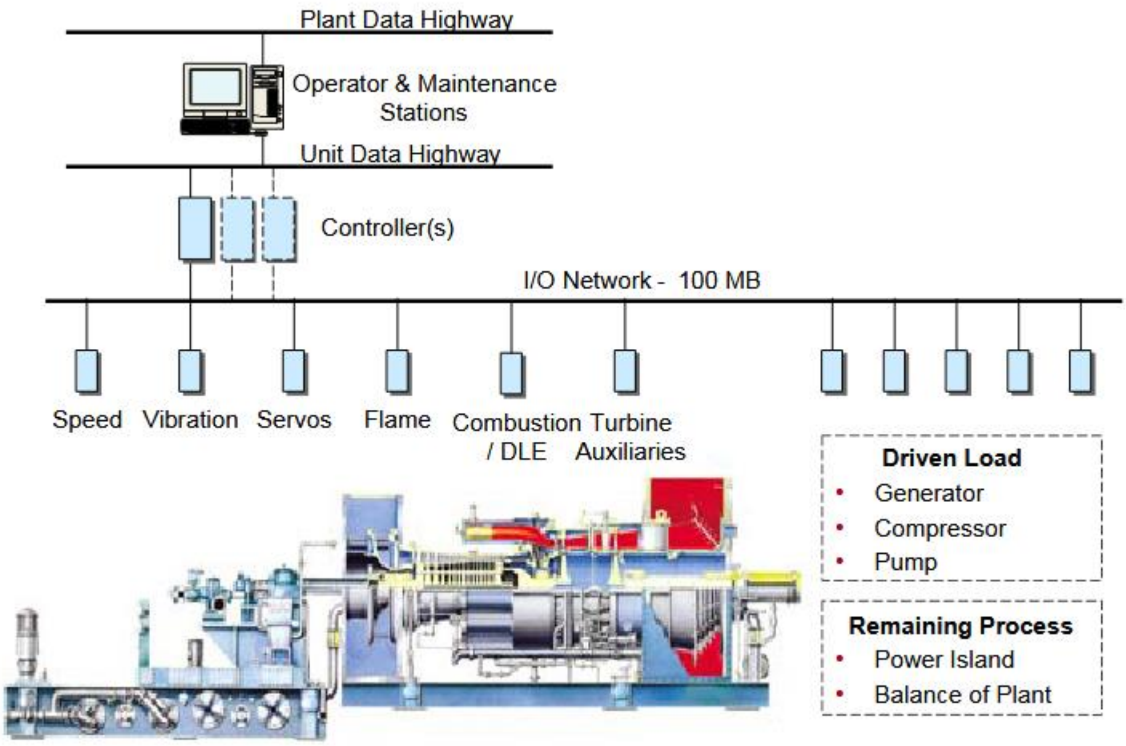

Figure 3.10: Architect system Mark VIe

.2 System Architecture Overview:

The Mark VIe system is built using a distributed control design, which means the control is not centered in one place. Instead, each control module works on its own but stays connected with the others through a fast Ethernet network. This setup makes the system more reliable, easier to expand, and quicker to respond in important situations.[5]

The main parts of the system are:

- Control modules (UCS or UCSB)
- I/O modules (Pack-based or VME)
- Communication interfaces
- Human-Machine Interfaces (HMIs) [5]

These parts are all connected using Ethernet networks that are separated based on their role: some handle control logic, others manage data, and some are used for HMI communication. This prevents delays in data flow and helps keep the system secure.[5]

There's also a network called IONet, which links the I/O packs to the controller. Other networks are used for HMI data and remote access. This way of separating the networks follows the latest standards for cybersecurity and real-time control.[5]

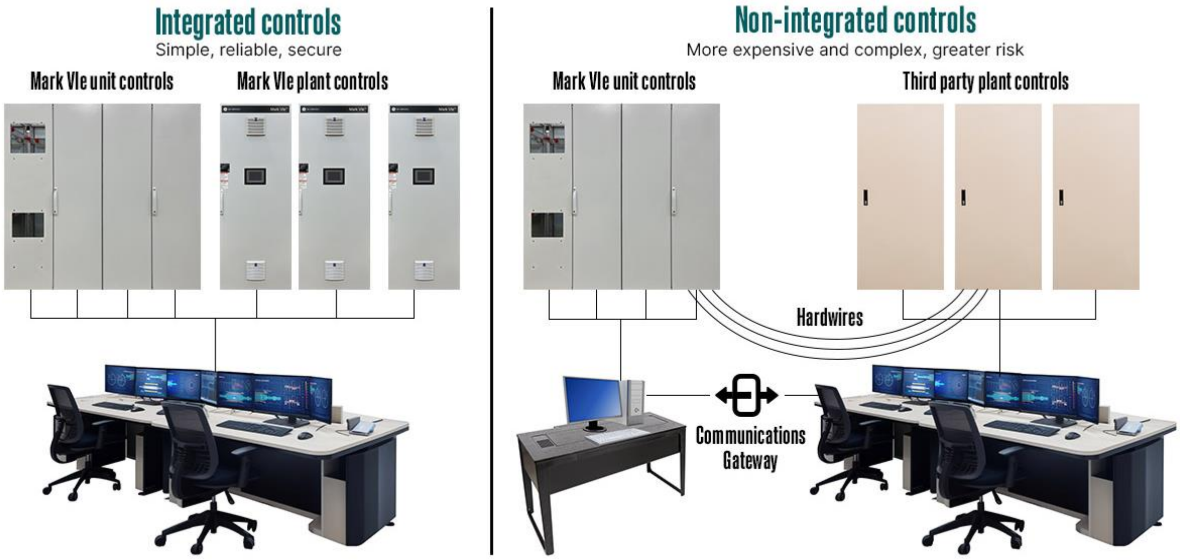

Figure 3.11: UCSC controllers linked via IONet to I/O packs and HMI/third-party systems. This architecture enables modular and distributed control across turbine and generator subsystems.

.3 Key Functionalities

The Mark VIe system from GE handles many important tasks that help industrial equipment run safely and efficiently. These tasks include:

- Real-time control and logic processing
- Collecting data from sensors and actuators
- Protection functions like overspeed, overtemperature, and overcurrent
- Managing start-up and shutdown steps
- Running PID control loops for different process values
- Handling alarms and events

- Doing diagnostics and trend analysis [6]

One of the key features of the system is that it runs control logic in a very precise way. The tasks are scheduled exactly on time, which helps avoid any delays. This makes the system reliable and predictable, especially in turbine operations where timing is critical.[6]

.4 SPEEDTRONIC series evolution

The Mark VIe is part of the SPEEDTRONIC family, which is a line of turbine control systems made by GE since the 1960s. Every new version in this series came with better control features, improved safety, and stronger hardware.[7]

Evolution of SPEEDTRONIC:

- **Mark I - Mark IV**: These were early systems that used analog and simple discrete logic for basic turbine control.
- **Mark V**: Was the first to use digital control with microprocessors. It had better logic functions and could do diagnostics.
- **Mark VI**: Came with a more advanced HMI, digital inputs/outputs, and a flexible system design.
- **Mark VIe**: Took things further by adding Ethernet communication, distributed control, and a modular design.[7]

The "e" in **Mark VIe** stands for “enhanced,” meaning that it brought several improvements over earlier versions, like:

- Built-in Ethernet IONet for distributed I/O
- Better cybersecurity options
- I/O packs that are redundant and can be replaced while the system is running
- Easy integration with ControlST and ToolboxST software [7]

One of the good things about the SPEEDTRONIC series is that newer systems can still work with parts of older ones in some setups. This helps when upgrading, because it means you don’t always have to replace everything.[7]

.5 Hardware Components

The Mark VIe system consists of several main hardware parts, each with its own job:[8]

Controller Modules:

- The UCS or UCSB units are responsible for running the control logic.
- They can be set up in redundant configurations to keep the system working even if one part fails.
- These modules have the processors, memory, and network connections needed for operation. [4]

I/O Modules:

- These connect directly to field devices like sensors, switches, and actuators.
- They come in different configurations: Simplex, Dual, or TMR (Triple Modular Redundancy) for higher reliability. [4]
- Examples of I/O modules include:
    - **TIC** for thermocouple inputs
    - **YIC** for discrete inputs
    - **AIC** for analog inputs
    - **AOC** for analog outputs

Power Supply & Rack Systems

- These components are usually mounted in VME or DIN-rail racks.
- They support redundant power supplies to make the system more reliable.

Operator Interface / HMI

- The HMI is usually a PC-based system running CIMPLICITY software or part of the ControlST suite.
- It gives operators access to live plant data, control screens, alarms, and trend displays. [8]

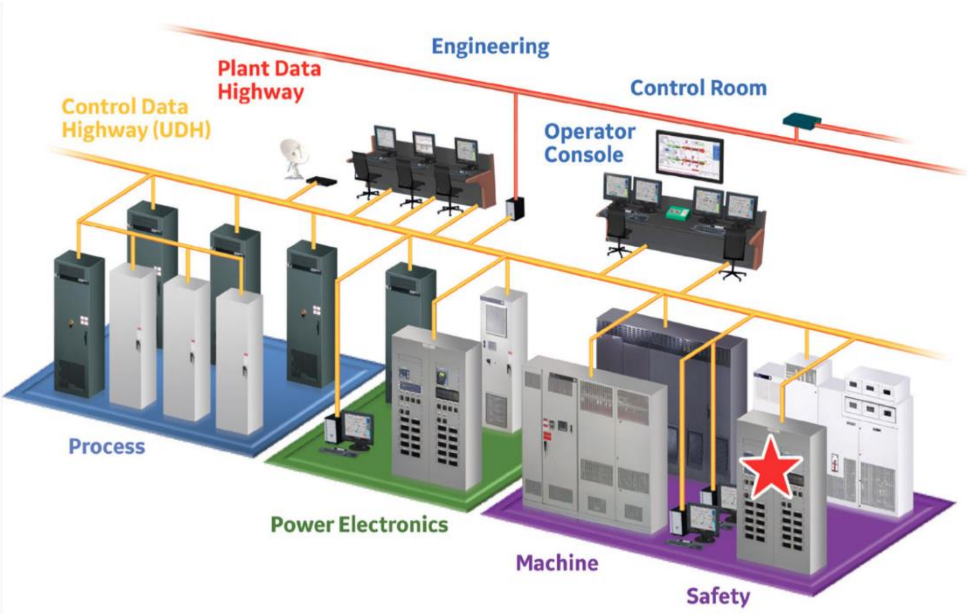

Figure 3.12: UCSC controllers, Ethernet communication, and modular I/O packs form the hardware backbone of the Mark VIe.

.6 Software Tools and Configuration Environment

The main software used with the Mark VIe system is **ToolboxST**, which is part of GE's ControlST suite. ToolboxST is an all in one tool that is used for:

- Creating control logic
- Configuring I/O settings
- Managing alarms
- Tracing signals and doing diagnostics
- Monitoring trends and system performance [9]

It uses block programming, where engineers build logic diagrams using standard or custom blocks. The software also allows live monitoring, which is helpful when checking for issues or analyzing the system while it's running.

Other tools in the ControlST suite include:

- **Workbench** – shows the system during runtime
- **Alarm Viewer** – tracks alarms and events
- **Device Configuration Tool** – used to set up hardware devices

ControlST also supports version control and audit trails, which are useful in industries like power generation and oil & gas where every change needs to be recorded. [9]

.7 Communication and Networking

The Mark VIe system uses several Ethernet-based networks to make sure communication stays fast and reliable:[10]

- **IONet**: Connects the I/O packs to the controllers, and can use either fiber-optic or copper Ethernet.
- **Control Network**: Manages communication between the controllers and the HMI.
- **Client Network**: Used for connecting with third party systems like SCADA or DCS.
- It also supports interfaces like **Modbus TCP/IP**, **OPC DA/UA**, and **Profibus** to work with other external systems.

Using both fiber and copper Ethernet helps reduce delays and protects the system from electrical noise (EMI). Time synchronization is also supported through **IEEE 1588 PTP** or **NTP**, which is important for accurate event recording and protection features. [10]

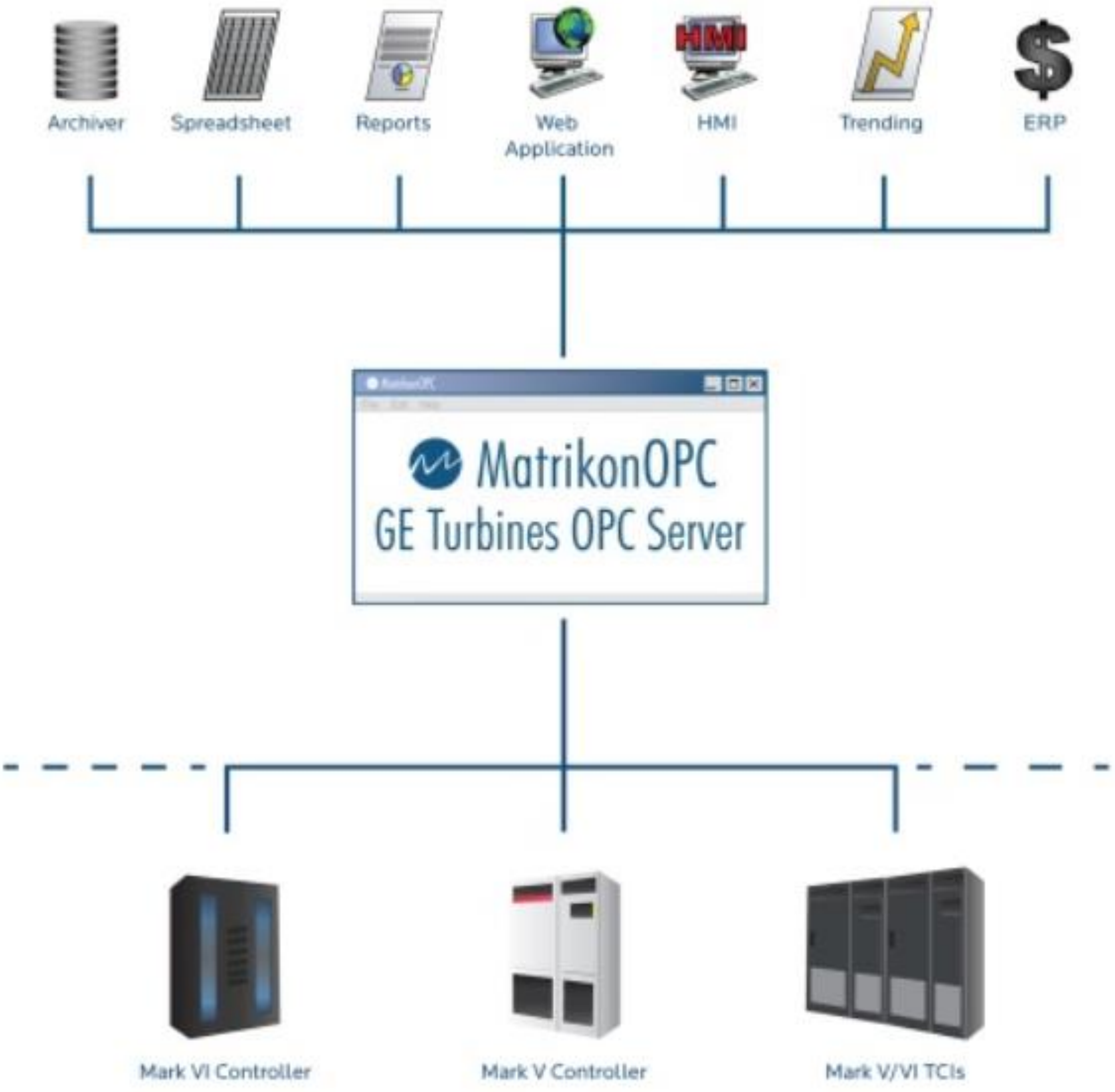

Figure 3.13: OPC server diagram showing bi-directional data exchange between Mark VIe controller and SCADA/DCS systems for event monitoring, alarm reporting, and process visualization.

.8 Industrial Applications

The Mark VIe system is used in many industries thanks to how reliable and flexible it is Some of its main applications include: [11]

- **Gas Turbines**: Controls things like combustion, speed, temperature, and provides protection functions.
- **Steam Turbines**: Manages valve movements, overspeed protection, and keeps the generator in sync.
- **Compressors**: Handles anti-surge control and supports different auxiliary systems.
- **Power Plants**: Often works as part of a larger Distributed Control System (DCS) to control the whole plant.

- **Oil & Gas Facilities**: Found in key areas like refineries and LNG units where reliable control is critical. [6]

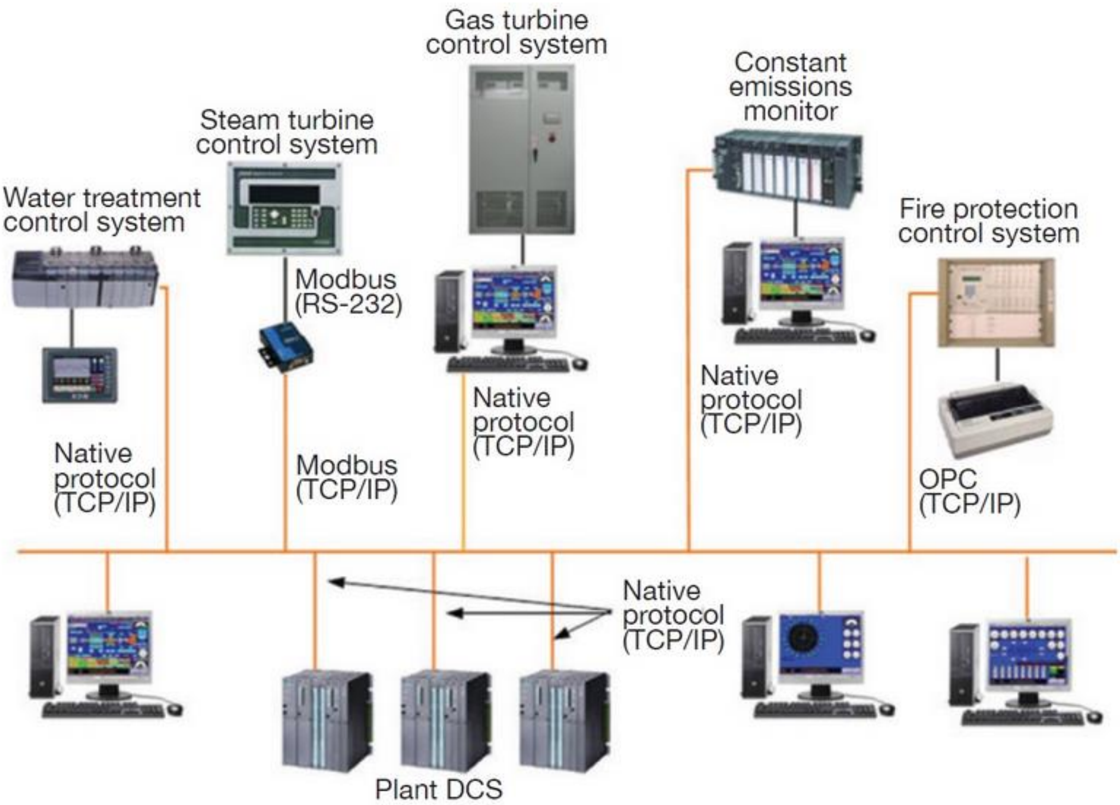

Figure 3.14: Mark VIe system interfaced with supervisory DCS and common HMI terminals using OPC/Modbus protocols.

.9 Advantages and Limitations

Like any advanced control system, the Mark VIe has its strengths and a few challenges. Here's a quick look at both:[12]

Advantages

- The system is modular and scalable, so it can be adapted to fit small or large applications.
- It offers high reliability, especially when using the TMR (Triple Modular Redundancy) setup.
- ToolboxST provides a complete set of tools for configuration, diagnostics, and monitoring.
- Supports remote diagnostics and secure access, which helps in maintenance and troubleshooting.
- It can communicate with many other systems thanks to its wide protocol support.

Limitations

- The system can be complex for people who are new to it.
- It uses proprietary components, which can increase costs.

- Specialized training is needed to work with it effectively.
- Cybersecurity updates are tied to GE’s licensing and support.

.10 Conclusion

The GE Mark VIe is a strong and flexible control system that fits well in today’s demanding industrial environments. Thanks to its modular structure, powerful communication features, and complete software tools, it has become a trusted choice in many industries.[13]

For engineers and researchers working on predictive maintenance, it’s important to understand how the Mark VIe handles data, I/O connections, and control logic. This knowledge is key to building smarter and more efficient systems.

Looking ahead, combining the Mark VIe with machine learning and predictive analytics opens the door to fewer unexpected failures and better performance in critical operations.[13]

Chapter 4: Machine learning and it's methods

.1 Introduction:

Since their evolution, humans have constantly used tools to simplify various tasks. The idea of intelligent machines has existed for centuries but has recently gained significant importance, particularly in the industrial and computing sectors, thanks to the advent of machine learning.

Machine learning, at the core of data science and artificial intelligence, plays a key role in the digital transformation of businesses. Its applications are diverse, now extending to the fields of mechanics and industrial maintenance.

At its foundation, machine learning revolves around data modeling, drawing on principles from statistics, artificial intelligence, and signal processing.

This chapter explores different machine learning approaches for analyzing data based on their specific characteristics. We will discuss the definition, types, and methods of machine learning, with a particular focus on its application in industrial maintenance to enhance equipment reliability and performance.

.2 Machine learning:

Machine Learning is a branch of artificial intelligence that enables computers to learn autonomously, without being explicitly programmed. To do so, machines rely on data for analysis and training. Big Data plays a central role in this process, serving both as the foundation for Machine Learning and as the resource it helps to fully exploit.

Machine Learning excels in extracting insights from large, varied, and constantly evolving data sets collectively known as Big Data. Compared to traditional analytical approaches, it offers greater speed and accuracy. For instance, by analyzing transaction details like amount and location, along with historical and social data, Machine Learning can identify potential fraud in milliseconds. This makes it far more efficient than conventional methods for processing transactional, social media, or CRM data.

However, the effectiveness of Machine Learning is still limited by the volume and quality of the input data. Analyzing something like a standard image, for example, involves processing thousands of pixels. To manage this, systems must be designed to receive, group, and filter relevant information for the algorithm. This is precisely where deep learning becomes essential.

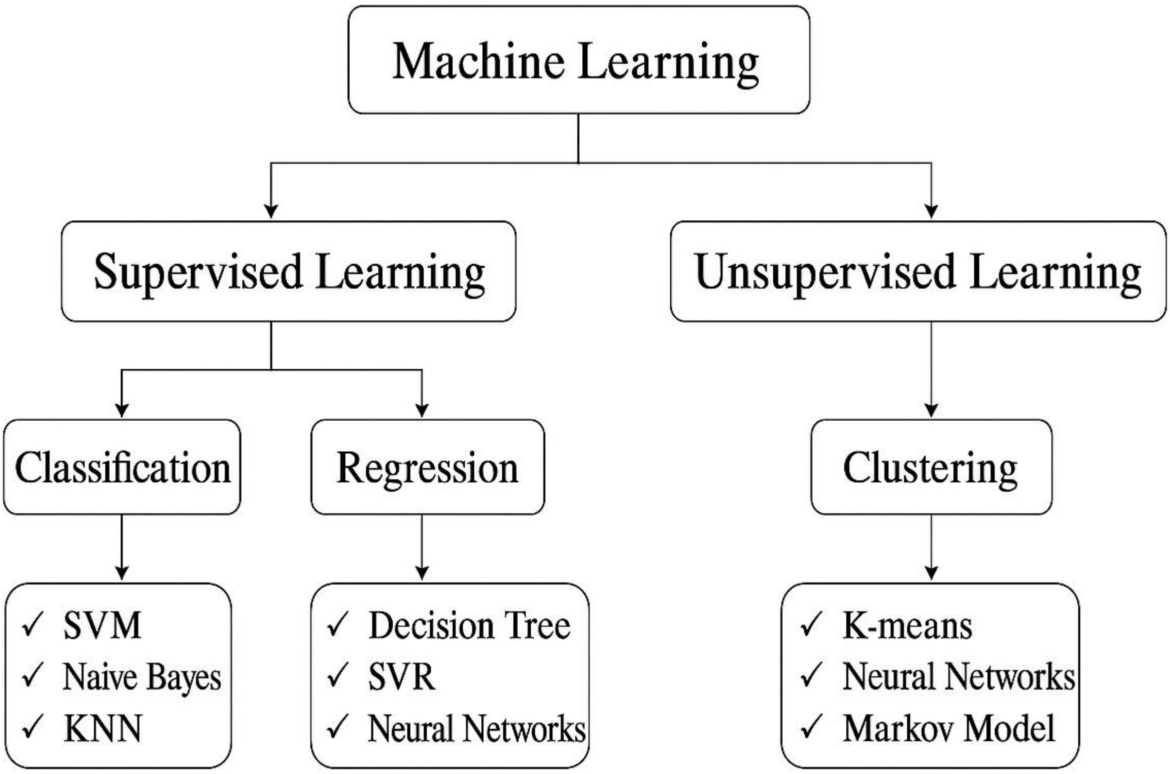

Figure 4.15: Taxonomy of Machine Learning

.3 Types of Machine Learning Algorithms:

a. Supervised Machine Learning Algorithms:

Supervised machine learning is the most widely used type. In this method, a data scientist guides the algorithm by providing it with labeled data and the correct outcomes it should learn to predict. Common algorithms in this category include classification, linear regression, logistic regression, and support vector machines.

Here, the algorithm is trained using input-output pairs. It learns patterns by comparing the input data to the known results. Once trained, it can apply what it learned to new, similar data. The system groups similar items together based on the training data and makes predictions. The analyst or researcher then interprets these results to form conclusions.

This process is similar to how a child learns to recognize fruits by studying labeled pictures in a book. The labels help the algorithm understand what to look for.[14]

In Figure 4.16, labeled data is used to train the machine learning model. The algorithm adjusts its parameters during training to create a logical prediction model.

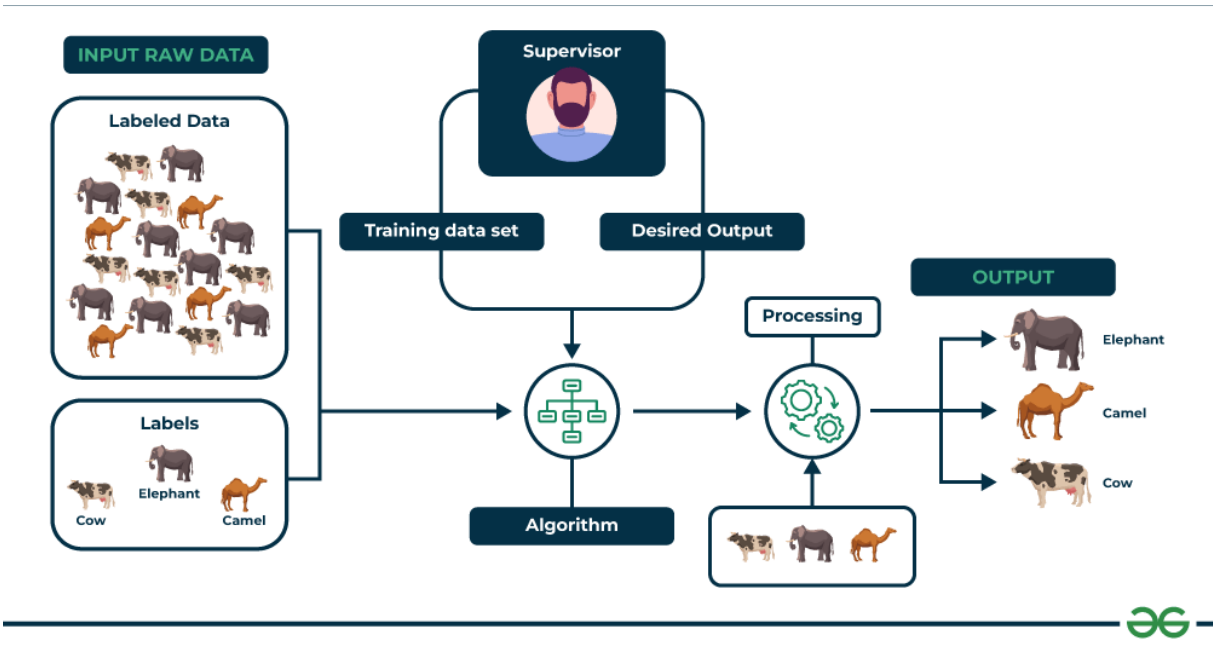

Figure 4.16: Supervised Learning

b. Unsupervised Machine Learning Algorithms:

Unsupervised machine learning takes a different path. It works with data that has no labels or predefined answers. The system learns to find patterns, groupings, or structures in the data on its own.[14]

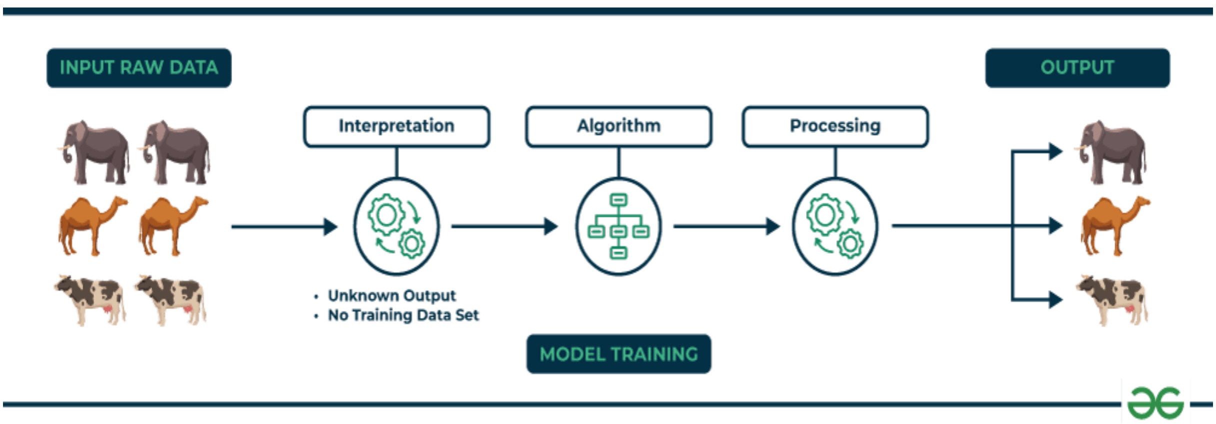

Figure 4.17: Unsupervised Machine Learning

Supervised learning methods are divided into two main categories based on the type of ML tasks they aim to solve

.4 CLASSIFICATION:

This form of supervised machine learning attempts to create rules for the sorting of objects with respect to the qualitative or quantitative features. It looks at certain features of the data and assesses how they ought to be defined or interpreted.[15]

Classification tasks engage in predicting categorical output labels or responses on the basis of what the model learns in the training phase. These output labels also called classes or class labels are categorical or discrete and unordered values. Each output belongs to a specific, distinct category.[16]

Multi-class classification results when the number of possible output classes is greater than two. In this scenario, every prediction can belong to one of the possible classes. The output in single class and multi-class cases is a scalar for a particular class.

The output in multi-label classification is often a vector containing one or multiple class labels for a given input sample. An example from the real world would be predicting the category of a news article which may be any category such as news, finance, politics, etc.

Some of the common classification algorithms are logistic regression, support vector machines, neural network, ensemble methods like random forests and gradient boosting, k-nearest neighbours, decision trees and many others.[16]

.5 REGRESSION:

Regression is used to predict a value from some input variables. One can say the above output data under this method associates, all input data is through a mathematical relationship the output data. Therefore, this means we can predict future data (not exact values but probable values) with the help of this model.[17]

Linear regression is one of the most popular regression approaches which attempts to establish a linear relationship between variables. Other types of regression include, logistic regression for predicting binary or categorical variables.[17]

The objective of regression is to minimize the difference between estimated and true values, enabling accurate predictions based on new data and the construction of dependable patterns.

.6 Machine Learning Within Industry 4.0

Industry 4.0 has reshaped traditional industrial systems into intelligent, connected, and data-centric environments. This shift relies on key technologies such as the Industrial Internet of Things (IIoT), cloud computing, big data, and artificial intelligence where machine learning plays a vital role.[18]

In this setting, machine learning provides predictive abilities that surpass rule-based automation. By processing large amounts of real-time data from sensors, controllers, and networks, it can recognize patterns, detect irregularities, and forecast issues before they occur.[19]

In systems like the GE MARK VIe, widely used for turbine and plant control, machine learning helps enable predictive maintenance and system health tracking. This directly supports Industry 4.0 goals: greater efficiency, less downtime, and smarter, data-driven decisions.[20]

Machine learning models built into control systems can learn from ongoing data, adjust to new conditions, and alert operators to early signs of equipment failure. These features are crucial to achieving the performance standards expected in Industry 4.0 environments.[20]

.7 Conclusion

Machine learning plays a vital role in Industry 4.0 by turning raw industrial data into useful insights. Its strength lies in learning from data, recognizing new patterns, and making accurate forecasts, making it well-suited for today’s smart industrial systems.[18]

Integrating machine learning into systems like the GE MARK VIe brings powerful predictive features such as early fault detection, live diagnostics, and better maintenance scheduling. These functions align with Industry 4.0 goals: smarter processes, less downtime, and greater system reliability.[20]

Merging machine learning with Industry 4.0 helps industries build more autonomous, efficient, and dependable systems ready to face future demands.[19]

Chapter 5: Methodology and results

.1 Introduction

This chapter outlines the steps taken to build and evaluate a predictive model based on industrial system data collected from the GE MARK VIe Speedtronic control system operating in Hassi R'mel. The aim is to improve prediction accuracy of a key temperature measurement (TTXD) using machine learning. The selected input variables are inlet pressure, inlet temperature, and IGV (Inlet Guide Vane) position. These features were chosen for their relevance to thermal dynamics and flow control in turbine operation. The study applies and compares three regression models Linear Regression, Random Forest, and Support Vector Regression using MAE, MSE, and R^2 as evaluation metrics.

.2 Data Collection and Filtering

The dataset used in this study was collected from the control systems of the GE Speedtronic MARK VIe, operating at the Hassi R’mel gas field. This industrial site is equipped with advanced instrumentation that records a wide range of operational parameters in real-time. These parameters are essential for monitoring, control, and optimization of turbine-based power generation processes.

To ensure the study remained focused and relevant, a feature selection process was applied to filter the raw data. From the full dataset, only the most influential variables were retained for model training. The selected input features were:

- Inlet Pressure
- Inlet Temperature
- IGV (Inlet Guide Vane) Position

These variables were chosen because they are known to directly influence turbine behavior and heat transfer, making them significant contributors to changes in the output variable.

The target variable selected for prediction was:

TTXD representing a critical temperature reading in the system, which is commonly used as a reference for thermal load and turbine condition.

By narrowing the dataset to these four variables, the study could focus on evaluating how inlet conditions and mechanical settings (IGV position) affect the thermal output (TTXD). This filtering step reduced noise in the data and improved model training efficiency while maintaining the physical relevance of the predictive task.

.3 Study Focus:

The focus of this study is to predict the value of a key system output, TTXD, based on three input variables: Inlet Pressure, IGV Position, and Inlet Temperature. These inputs were chosen because they directly affect turbine operation and heat transfer inside the system.

By studying the relationship between these inputs and the TTXD output, the goal is to develop a reliable predictive model. This model can help operators better understand system behavior and detect potential issues early.

To achieve this, three machine learning algorithms were used and compared:

Linear Regression, to test a simple linear relationship.

Random Forest, to capture more complex patterns using multiple decision trees.

Support Vector Regression (SVR), to model non-linear relationships using a kernel approach.

The comparison helps determine which method offers the most accurate and consistent predictions for real industrial use.

.4 Data Preprocessing

To train and evaluate the models, the dataset was split into training (80%) and testing (20%) subsets using a fixed random seed to ensure consistency across experiments.

No additional feature scaling or normalization was applied because each algorithm internally handles data in a way that is appropriate for its structure. For instance, Random Forest is insensitive to feature scaling, while SVR with an RBF kernel can adapt to unscaled input if the data is moderately balanced.

Since the output variable is continuous, regression models were appropriate for this predictive task.

.5 Machine Learning Models

After preparing the dataset, three machine learning models were developed to predict the TTXD output. The models used were:

.5.1 Linear Regression (LR)

Linear Regression assumes a linear relationship between the input features and the target variable. It serves as a baseline model and helps determine how well a simple linear approach can capture patterns in the data.

.5.2 Random Forest Regression (RF)

Random Forest is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions. It is robust to overfitting and can model complex non-linear relationships. For this study, the model was configured with 10 decision trees and a maximum tree depth of 2, which limits model complexity and controls overfitting.

.5.3 Support Vector Regression (SVR)

SVR is a powerful regression method based on support vector machines. It is especially effective in modeling non-linear relationships and handling noise. The model used an RBF (Radial Basis Function) kernel with the following hyperparameters:

- C = 100: Penalty parameter that controls the trade-off between smoothness and accuracy.
- gamma = 0.1: Determines the influence of a single training example.
- epsilon = 0.1: Defines a margin of tolerance for error in predictions.

Each model was trained using **80% of the data** and tested on the remaining **20%**. This training/test split helps evaluate how well each model performs on unseen data.

The models were implemented using Python's **scikit-learn** library. The training process involved fitting the models to the input features (Inlet Pressure, IGV Position, and Inlet Temperature) and learning how they relate to the TTXD output.

To measure how accurate each model was, three standard error metrics were used:

- **Mean Absolute Error (MAE):** Measures the average size of prediction errors.
- **Mean Squared Error (MSE):** Penalizes large errors more heavily.
- **R-squared (R^2):** Shows how much of the TTXD variation is explained by the model.

These metrics were calculated on both training and test results for each model. This helps identify which model makes the best predictions and whether any model is overfitting or underfitting the data.

.6 Results Summary

The results, including predicted values, residual distributions, and metric scores, are provided in the Results chapter. This methodology ensures a balanced and reproducible evaluation of predictive accuracy using real-world industrial data.

A- **Predicition resluts**

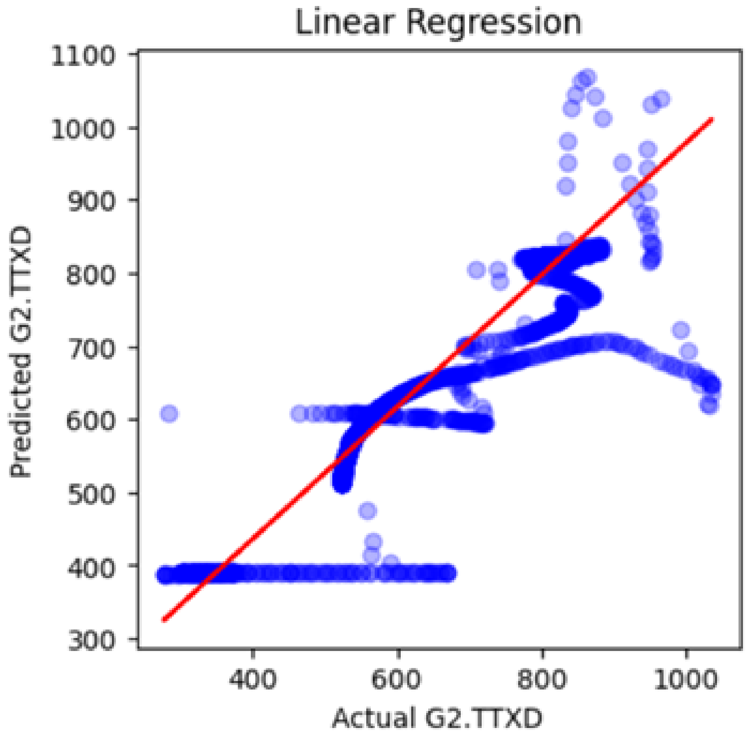

Figure 5.18: Linear Regression

The scatter plot in Figure 5.18 illustrates the performance of the Linear Regression model in predicting the G2.TTXD output variable. The horizontal axis represents the actual observed values, while the vertical axis shows the corresponding values predicted by the model. The red diagonal line indicates the ideal case where predicted values match the actual values exactly (i.e., the line of perfect prediction).

The distribution of data points around the diagonal line reveals several important characteristics of the model's performance. While a general positive correlation is observable, significant deviations from the ideal line are present. The predictions tend to cluster at specific output levels most notably around 400 and 600 which indicates that the model fails to fully capture the variability in the data. This behavior suggests a tendency toward underfitting, where the model is overly simplistic and unable to adapt to the complexity or non-linearity of the input-output relationship.

Furthermore, in the higher range of actual G2.TTXD values (e.g., above 800), the model systematically underestimates the output. Conversely, some degree of overestimation appears in the mid-range values. These prediction errors point to the limitations of linear regression in capturing non-linear patterns within the data, particularly in industrial systems characterized by dynamic and non-linear behavior.

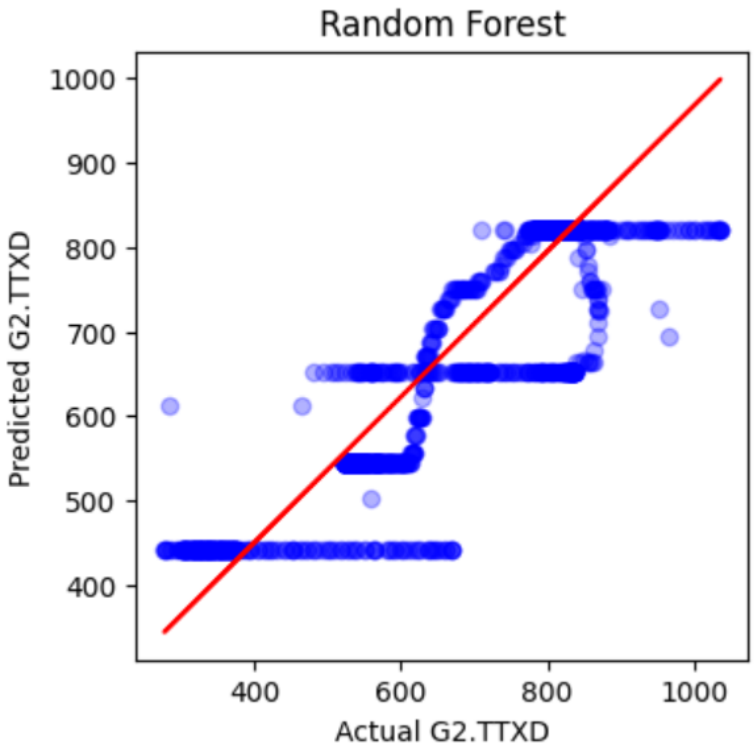

Figure 5.19: Random Forest

This plot presents the prediction results of the **Random Forest Regression** model. The ideal line of perfect prediction is shown in red.

The model captures general trends in the data but produces **step-like prediction patterns**, a common characteristic of tree-based models with limited depth or estimators. The predicted values appear to cluster at fixed intervals, especially around 400, 500, 600, 700, and 800, rather than forming a continuous prediction curve. This staircase effect indicates that the model's output resolution is constrained by the number and depth of trees.

Despite this, the Random Forest model still manages to follow the overall upward trend of the actual values. However, several regions especially at higher TTXD values show a **noticeable deviation from the ideal line**, suggesting a reduction in prediction accuracy at the extremes.

In summary, while Random Forest demonstrates reasonable performance, especially in mid-range values, it lacks the fine-grained prediction capability of SVR. Its performance could be improved by increasing the depth and number of trees to reduce output discretization and better capture subtle variations in the data.

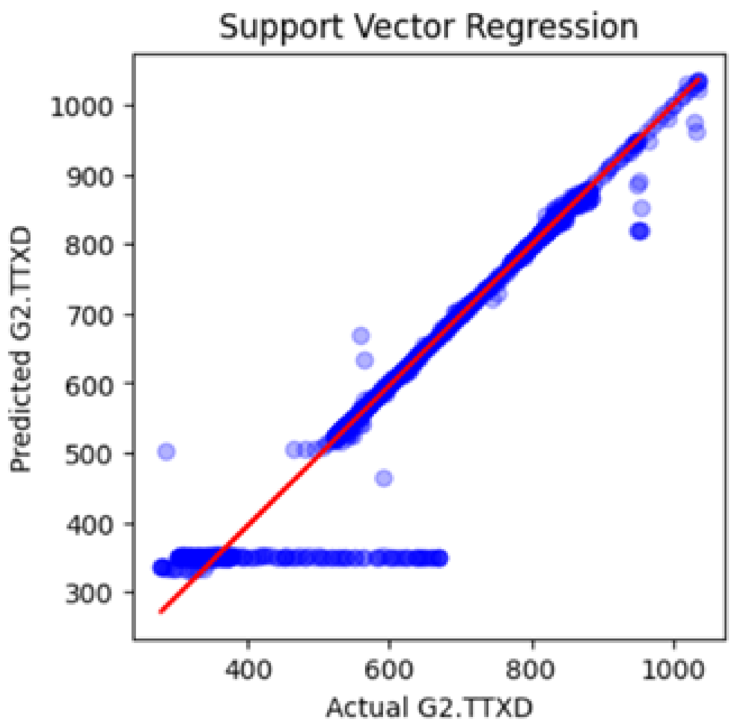

Figure 5.20: Support Vector Regression

This plot illustrates the performance of the **Support Vector Regression (SVR)** model in predicting G2.TTXD. The red diagonal line represents the ideal prediction scenario where the predicted values are equal to the actual values.

The distribution of data points shows a **strong alignment along the red line**, indicating that SVR captures the relationship between the input features and TTXD more accurately than Linear Regression and Random Forest. Most predictions closely follow the ideal trajectory, with only slight deviations observed across the value range.

However, a few instances of clustering around 300 - 400 on the predicted axis suggest that the model may compress predictions under certain conditions, possibly due to limited data diversity in those ranges or SVR's inherent regularization behavior.

Overall, the SVR model demonstrates **superior predictive accuracy** and generalization capability, particularly in handling non-linear patterns in the dataset. The tight distribution of residuals and minimal outliers further confirm SVR's suitability for modeling the behavior of the GE MARK VIe system output.

B- Residual distributions:

RF Residual distributions:

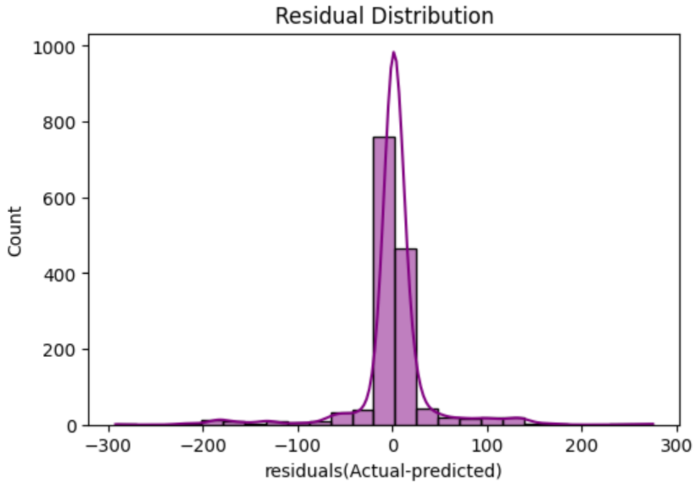

Figure 5.21:

Figure 5.21 illustrates the residual distribution for the Random Forest regression model, where residuals are defined as the difference between actual and predicted values (Actual Predicted). A well-performing regression model should produce residuals that are symmetrically distributed around zero, ideally forming a normal distribution. This indicates that the model does not systematically overpredict or underpredict the target values.

In this plot, the residuals are centered around zero, with a high concentration of values within a narrow range (approximately ±50). This indicates that the model has relatively low prediction error for the majority of the data points. However, the presence of long tails extending on both sides of the distribution suggests the existence of some outliers or larger errors in certain predictions. This behavior is characteristic of tree-based models, which may overfit or underfit in certain regions depending on the depth and number of trees used.

The sharp central peak also implies that the Random Forest model performs consistently well for most data points, but the kurtosis (peakedness) hints at limited flexibility in modeling continuous changes, likely due to its discrete, step-wise prediction nature.

Overall, while Random Forest exhibits good general performance with most residuals close to zero, further tuning of hyperparameters (e.g., tree depth, number of estimators) could help reduce the error spread and improve model robustness**.**

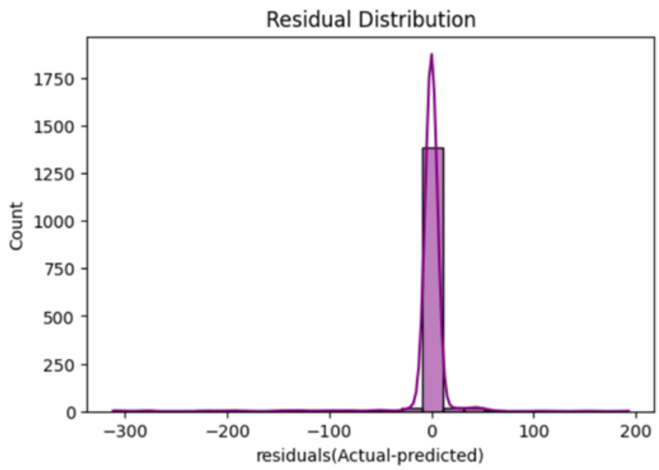

Figure 5.22: Svr Residua

Figure 5.22 presents the residual distribution for the Support Vector Regression (SVR) model, showing the difference between actual and predicted values (Actual Predicted). A desirable residual distribution is one that is centered around zero and approximately symmetrical, suggesting that the model does not systematically over- or under-predict.

In this case, the residuals are tightly concentrated around zero, with an extremely narrow spread and a sharp central peak. This pattern indicates that the SVR model achieves high prediction accuracy across most of the dataset. The minimal spread of residuals demonstrates the model's ability to generalize well without introducing large errors.

Moreover, the shape of the distribution is close to a normal distribution, reinforcing the assumption that prediction errors are random and not influenced by any strong bias or systematic pattern. Only a few outliers are observed on the tails of the distribution, further confirming the model's robustness.

This outcome aligns with the performance metrics observed earlier (e.g., low MSE and high R^2), and suggests that SVR is well-suited for modeling the complex, non-linear relationships in the input data of the GE MARK VIe system.

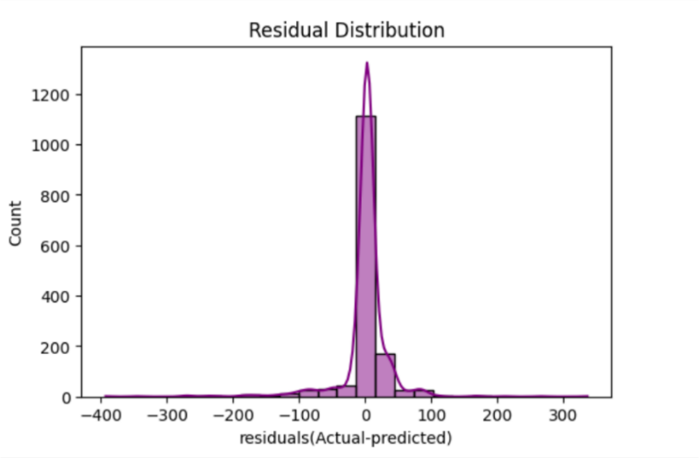

Figure 5.23: Predicition resluts

Figure 5.23 shows the residual distribution for the Linear Regression (LR) model, where residuals represent the difference between actual and predicted values (Actual Predicted). A good regression model typically yields residuals that are symmetrically distributed around zero with minimal spread, indicating accurate and unbiased predictions.

In this case, the residuals are **concentrated near zero**, suggesting that the model captures the general direction of the data. However, the **distribution is notably more skewed and dispersed** compared to that of SVR or Random Forest. The tail on the right side indicates the presence of **systematic under-predictions**, while the long left tail shows the model occasionally **over-predicts** by a significant margin.

The distribution is relatively peaked but also displays **moderate kurtosis and skewness**, suggesting **higher variance** and **less stable prediction behavior**. These patterns imply that the Linear Regression model lacks the flexibility to accurately capture the complex, likely non-linear, relationships present in the dataset.

In summary, while Linear Regression performs reasonably well in the central range of the data, the broader spread and asymmetry of its residuals point to **limited suitability** for this application. This reinforces the conclusion that **more advanced, non-linear models** such as SVR are better suited for predicting TTXD in the GE MARK VIe system context.

Value evolution :

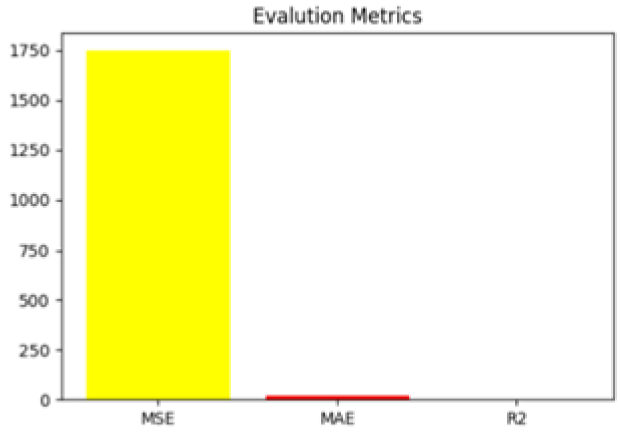

Figure 5.24: Random forest

Figure 5.25 illustrates the evolution of predicted and actual values of G2. TTXD over the sample index using the Random Forest Regression model. The plot compares how well the model tracks changes in the actual values across the sequence of observations.

The general trend in the predicted values follows the actual TTXD curve, indicating that the model has learned the broad patterns in the data. However, the predicted line displays a step-wise progression, which is characteristic of tree-based models like Random Forest. This behavior occurs because the model partitions the feature space into discrete regions, leading to piecewise constant predictions rather than smooth transitions.

Although the model succeeds in capturing general upward and downward trends, it struggles with precise alignment during rapid or subtle fluctuations in the actual values. As a result, the prediction curve appears less fluid and may lag or overshoot in regions where TTXD changes rapidly.

This visualization confirms that Random Forest is capable of providing a reasonable approximation of the system behavior but lacks the fine resolution needed for highly dynamic response tracking. Enhancing model granularity could improve its responsiveness to finer variations in the TTXD signal.

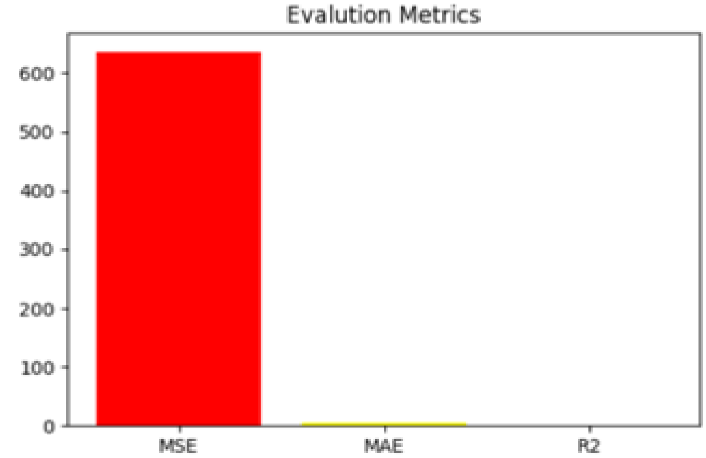

Figure 5.26: Support VR

Figure 5.25 presents the value evolution of the **G2.TTXD** variable, comparing the actual values with those predicted by the **Support Vector Regression (SVR)** model across sequential samples. This figure is used to assess how well the model follows real system behavior over time.

The SVR prediction curve closely mirrors the actual values throughout the entire sample range. This tight alignment reflects the model's strong generalization ability and its effectiveness in capturing both the trend and magnitude of TTXD variation. SVR's ability to learn non-linear relationships through kernel functions such as the radial basis function (RBF) allows it to respond to fluctuations in the system output with high precision.

Notably, the predicted values exhibit **smooth and continuous transitions**, without the abrupt jumps seen in tree-based models. This smoothness suggests that the SVR model provides **stable and high-resolution forecasts**, which is particularly valuable in control and monitoring systems like GE MARK VIe, where fine-scale prediction supports early fault detection and real-time optimization.

In conclusion, the SVR model demonstrates the most consistent and accurate value evolution among the tested models, reinforcing its suitability for predictive tasks in industrial automation contexts requiring high fidelity.

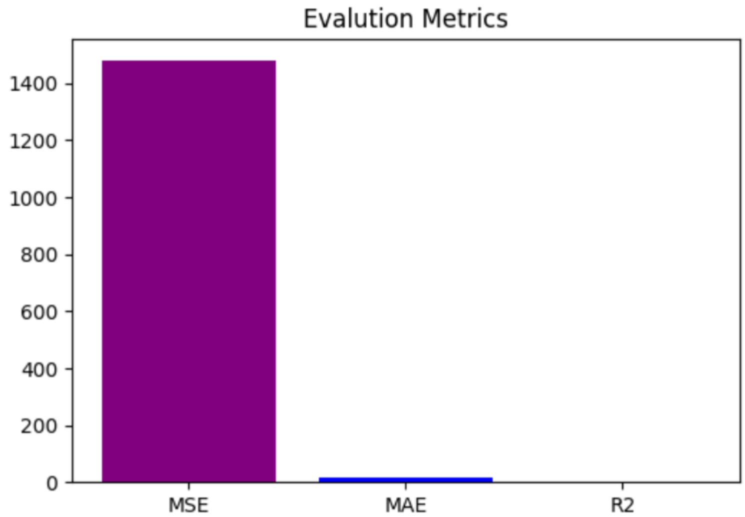

Figure 5.27: Linear regression

Figure 5.26 illustrates the evolution of the predicted and actual values of GQ2. TTXD over time or sample index, as modeled by the Linear Regression algorithm. The visualization compares the model's output with the true system values, offering insight into its prediction consistency and temporal tracking ability.

While the Linear Regression model captures the overall direction of the trend, its predictions consistently deviate from the actual values, particularly during periods of rapid change or non-linear behavior. This is expected, as Linear Regression assumes a constant linear relationship between inputs and output, making it less effective in systems governed by complex dynamics, such as turbine control environments.

The predicted curve appears flattened and less reactive to fluctuations in the actual data. This leads to underestimation in rising segments and overestimation in falling segments. These distortions result in a notable gap between the prediction and observation lines, especially in areas with high variability in TTXD.

In summary, while Linear Regression provides a basic understanding of the system's direction, it lacks the sensitivity and flexibility needed for accurate value tracking in real-time industrial applications. This outcome highlights the importance of using more adaptive models such as SVR or ensemble methods for predictive tasks in non-linear control systems like the GE MARK VIe.

.7 Observations from Model Comparison

Tableau 5.2: Observations from Model Comparison

Method	Training MSE(°C²)	Training MAE(°C)	Training R^2 (%)	TEST MSE(°C²)	TEST MAE(°C)	TEST R^2 (%)
Linear Regression	1576.95	17.18	90.06%	1479.47	17.16	91.35%
Random Forest	1625.03	17.61	89.58%	1749.68	17.86	89.77%
Support VR	526.85	3.71	96.68%	635.77	4.96	96.28%

Table 5.1 provides a detailed comparison of the three regression models Linear Regression, Random Forest, and Support Vector Regression (SVR) applied in this study to predict the target variable G2. TTXD, based on selected operational features from the GE MARK VIe control system. The evaluation metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE), expressed in degrees Celsius (°C² and °C, respectively), along with the R-squared coefficient (R^2) represented as a percentage of variance explained by the model.

The results demonstrate that the SVR model consistently outperforms both Random Forest and Linear Regression in terms of accuracy and generalization. It yields the lowest MSE and MAE values for both training and test datasets, indicating minimal prediction error.

Furthermore, the SVR model achieves the highest R^2 scores 96.68% on the training set and 96.28% on the test set confirming its ability to capture complex non-linear relationships within the system data.

In contrast, the Random Forest model, while exhibiting reasonable accuracy, shows slightly higher error values and a wider gap between training and testing performance, suggesting mild overfitting. Its predictive performance is nonetheless superior to that of Linear Regression.

The Linear Regression model recorded the highest MSE and MAE values, indicating limited suitability for modeling the dynamic, non-linear behavior of the GE MARK VIe system. Despite achieving R^2 values above 90%, the model's relatively poor error performance reinforces the conclusion that linear methods are inadequate for capturing the intricacies of turbine system behavior in real-world operational settings.

These findings validate the choice of SVR as the most effective regression approach for predictive modeling in the context of this case study, aligning with the broader objectives of Industry 4.0 through improved accuracy, system awareness, and data-driven automation.

Based on your full set of results evaluation metrics, residual distributions, and value evolution plots the **Support Vector Regression (SVR)** model is clearly the **best choice for your machine learning approach** in this case study. Here's a summary of **why SVR is the superior method** for your GE MARK VIe predictive system:

Highest Predictive Accuracy

- **Lowest Test MSE**: 635.77 $°C^2$
- **Lowest Test MAE**: 4.96 °C
- These values indicate that SVR produces the **smallest average prediction error** compared to Random Forest and Linear Regression.

Best Generalization

- **Test R^2 Score**: 96.28%
- This high R^2 shows that SVR explains the largest proportion of the variance in the target variable (G2. TTXD), and performs well on unseen data.

Residual Distribution

- SVR's residuals are **tightly centered around zero** and show a near-normal distribution with very few outliers, meaning the errors are small, balanced, and not biased in one direction.

Value Evolution Curve

- SVR most closely **follows the actual output signal**, showing smooth, continuous predictions with minimal lag or deviation ideal for real-time system monitoring and predictive maintenance.

.7 Conclusion

This chapter presented the results of implementing and evaluating three regression models **Linear Regression**, **Random Forest**, and **Support Vector Regression (SVR)** to predict the key system output **G2.TTXD** based on selected operational parameters: Inlet Pressure, Inlet Temperature, and IGV Position.

through a comprehensive comparison of **performance metrics** (MSE, MAE, R^2), **residual distributions**, and **value evolution curves**, the **SVR model demonstrated clear superiority**. It consistently achieved the lowest error values and the highest explanatory power, with a **test R^2 of 96.28%**, indicating a strong ability to generalize to unseen data. Its residuals were narrowly distributed around zero, and its predicted values closely followed the real system behavior over time both essential characteristics for deployment in industrial predictive systems.

In contrast, **Random Forest** showed moderate accuracy but with noticeable step-like predictions and slightly higher variance in residuals. **Linear Regression**, while offering some insight into linear trends, was limited in handling the complex, non-linear relationships present in the system data, as evidenced by higher error values and less accurate tracking in time-based prediction.

In conclusion, **Support Vector Regression** is identified as the most effective and reliable method for predictive modeling in this context. Its ability to model non-linearity with high precision makes it highly suitable for **Industry 4.0 applications**, such as predictive maintenance and real-time diagnostics within the GE MARK VIe control environment.

References

[1] **SONATRACH Documentation:** Hassi R'mel Field

[2] **Manufacturer's Manual:** Northern Compression and Reinjection Station

[3] **NUOVO PIGNONE - Course:** MS 5002B Gas Turbines – Training Center

[4] **GE, GEH-6421L Mark VI Control System Guide, Volume I**: Overview and Operation, General Electric Company, 2018.

[5] **GE, GEH-6721A Mark VIe Control System Guide, Volume I:** System Overview and Architecture, General Electric Company, 2020.

[6] **GE, GEH-6721A Mark VIe Control System Guide, Volume I:** Functional Description, General Electric Company, 2020.

[7] **GE, GEH-6721, Volume III**: Mark VIe Control Systems Technical Manual – Hardware Description, General Electric Company, 2020.

[8] **GE, GEH-6721, Volume I:** Mark VIe Control System Guide – Reliability and Maintenance, General Electric Company, 2020.

[9] **GE Vernova,** Mark VIe Applications in Power Generation and Oil & Gas, GE Vernova Plant Controls Division, 2021.

[10] **GE Vernova,** Mark VIe Integrated Plant Control – ControlST Software Suite Overview, GE Vernova, 2022.

[11] **GE Vernova,** Mark VIe Control System Communication & Networking Overview, GE Vernova, 2022.

[12] **GE Vernova,** Lifecycle & Predictive Maintenance Enhancements in Mark VIe Systems, GE Vernova, 2023.

[13] **Kolotcholoma,** Support Introduction au Contrôle Commande Mark V & Mark VIe, training document, SONATRACH Training Center, Hassi R'mel, 2021.

[15] **J. W. Hanz, Machine Learning:** Algorithms, Real-World Applications and Research Directions. Springer, 2021.

[16] M. Rahaman, M. Islam, and R. Singh, "A Comprehensive Review to Understand the Definitions, Types, and Applications of Machine Learning Algorithms," ResearchGate, 2023.

[17] V. Saeed, M. M. Shahriari, and A. Khosravi, "A Brief Review of Regression-Based and Machine Learning Methods for Industrial Prediction Tasks," PMC, 2012.

[18] M. Achouch, L. Bougueroua, and N. Zerhouni, "On Predictive Maintenance in Industry 4.0: Overview, Models, and Challenges," Applied Sciences, vol. 12, no. 8, pp. 4001–4023, 2022.

[19] M. Vulić, T. Jurišić, and B. Vrdoljak, “Machine Learning for Predictive Maintenance in Industry 4.0: A Systematic Literature Review,” Acta Technica Corviniensis, vol. 17, no. 1, pp. 45–50, 2024.

[20] R. B. Gopaluni, B. Huang, and A. D. Mesbah, “Modern Machine Learning Tools for Monitoring and Control of Industrial Processes,” arXiv preprint, arXiv:2210.01776, 2022.