ISSN (O) 2321-2004, ISSN (P) 2321-5526

IJIREEICE

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 9, Issue 12, December 2021

DOI: 10.17148/IJIREEICE.2021.91215

Integrating Machine Learning and Data Engineering for Predictive Maintenance in Smart Agricultural Machinery

Sathya Kannan

Sr AI Developer, ORCID ID : 0009-0009-1010-2493

Abstract: Agriculture plays a significant role in the global economy, and there has been a trend in industrializing agriculture machinery and equipment. This paper proposes a framework that integrates data engineering and machine learning for the predictive maintenance of smart agriculture machinery. This framework is built upon existing state-of-the-art solutions, concepts, and techniques for data engineering and machine learning, along with the innovations of new solutions. This paper highlights the major elements of the collaboration. Challenges, societal impact, and future works are further discussed.

The agriculture sector is a key contributor to the global economy, with an estimated total value of more than 3 trillion dollars per year and corresponding employment of over 1 billion people worldwide. The rapid growth of population increases the demand for basic agricultural supplies. In the era of IoT, pollution prevention, and food safety assurance, agriculture has also been trending towards intelligence, automation, and standardization. There has been a trend in industrializing agriculture machinery and equipment, where a wide range of data sources from working conditions are equipped on agricultural machinery. These data sources may include 1D, 2D, 3D, and 4D data from cameras, LIDARs, radars, and sensor networks. These data sources by design have the potential of being a bridge to connect agriculture and smart city. However, the wide-scale deployments of data-driven smart agriculture have been hindered by the challenge of data engineering for the large-scale, heterogeneous, and sparsely-distributed agriculture data, and insufficient integration of data exploitation and exploration technologies including machine learning for deep analysis, insight mining, and knowledge discovery of agriculture data.

Moreover, a variety of application scenarios in smart agriculture have appeared in recent years, including but not limited to predictive maintenance of agriculture machinery, soil monitoring with sensor networks, and enviromonitoring with remote sensing data for crop estimation. The laboratory research has made promising achievements in devising precise models with techniques from data analytics, machine learning, artificial intelligence, computer vision, and other similar fields. However, these achievements are hardly used in practice because of integration challenges. Integrating data engineering and machine learning for the collaborative, collaborative-wise, and process-wise predictive maintenance of smart agriculture machinery is yet to be studied comprehensively and in-depth.

Keywords: Predictive Maintenance,Smart Agriculture,Machine Learning,Data Engineering,IoT Sensors,Time Series Analysis,Remote Monitoring,Failure Prediction,Remaining Useful Life (RUL),Condition-Based Maintenance,Edge Computing,Big Data Analytics,Sensor Fusion,Agricultural Machinery,Maintenance Optimization.

1. INTRODUCTION

As the world enters a new era of economic competition, the adoption of advanced technological solutions as a means for the improvement of agricultural productivity and sustainability, among others, is accelerating. However, in the case of direct agricultural machines, such as soil tillage, seeding, fertilization, and plant protection, there is no relevant commercial technology and little research exists, despite the fact that such machines are multi-million euro investments per unit. In this context, the integration of advanced computational devices, such as data-logger, mesh network, IoT devices, and edge/yards, is converted into a Smart Farm IoT Architectures. Modern data collection technologies promote equipment Condition Monitoring and new opportunities for data management. Modern data engineering technologies are leveraged in cloud solutions, reducing the SMEs investments and maintenance of dedicated infrastructures providing the necessary tools for Data Storage and Processing. Edge computing devices mitigate the data-delivery bottlenecks of cloud architectures while safeguarding processing in case of intermittent Internet connections, not all data need to be delivered, and privacy concerns. The whole of the aforementioned technologies is complemented by Machine Learning technologies mostly aimed at Real Time Predictive Maintenance and their application to agricultural machinery. Conditions such as the malfunctioning of a component are considered. At this failure mode, maintenance schedules are shown to be unnecessary and are thus skipped. Future work will focus on abnormal conditions resulting in a lower tractability of the driller, or issues with undue clogging of seed. The deliverable is thus expected to be for a specific sensor type, a collection of pre-processing pipelines or data restructuring procedures producing an enriched timeseries scoping on the final use-case. Hence, for agricultural data analytics, these latter complications specifically require the design of a Data Engineering workflow, affected by raw data variability on both modelling and real-world terms of applicability (e.g. sensor noise, data dropouts, fuzzy input values). In line with these considerations, Machine Learning technologies are expected to be robust to noise. However, robust generalization is not necessarily achieved with expressive model families operating on raw input.

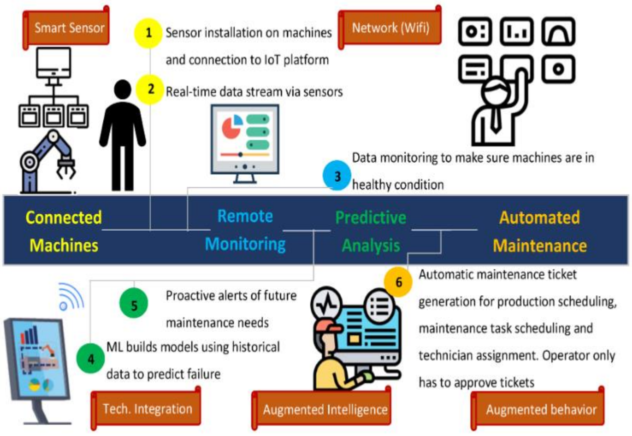
Smart Sensor
1 Sensor installation on machines and connection to IoT platform
Network (Wifi)
2 Real-time data stream via sensors
3 Data monitoring to make sure machines are in healthy condition
Connected Machines
Remote Monitoring
Predictive Analysis
Automated Maintenance
5 Proactive alerts of future maintenance needs
4 ML builds models using historical data to predict failure
6 Automatic maintenance ticket generation for production scheduling, maintenance task scheduling and technician assignment. Operator only has to approve tickets
Tech. Integration
Augmented Intelligence
Augmented behavior

Fig 1: Machine Learning and Data Engineering for Predictive Maintenance in Smart Agricultural Machinery.

1.1. Bridging AI and Agriculture: A New Era of Predictive Maintenance in Smart Machinery

In recent years, a revolution has begun to unfold in cultures around the world and it appears poised to change profoundly the agriculture industry. Technology has advanced to remarkable levels, computers, machine data-central engineering, data science, sensors, smart phones, robots, drones, and smart machinery cannot be overwhelming. Surplus labor of rural areas, dense population of urban area and starvation for food of a growing population in mid developing countries turns research attention to agri-robotics. This is a focus on a class of smart machineries capable of using smart agriculture. The ultimate goal is to develop self-operate machines capable of planting, growing, nurturing, and harvesting crops all on their own. These machines will be steered by artificial intelligence (AI). Fallible data will be collected on board by sensors for learning purposes on cultivation, nutrition, harvesting, weed and pest controlling, environmental or weather influence. It is then uploaded to the cloud server for consideration by compositing, analysing, studying, and learning. Precedent will develop an AI technology executed on-board for smart agriculture by the intelligent machines or it will relay results to the smart machines for actions. The smart machineries with smart soft wares are capable of autonomously performing data driven precision farming on lush crops and fruit trees. Along with such revolutionary research, there has been a recognized shift of understanding about preciseness in agriculture: agriculture engineering cannot only apply engineering principles and technology of manufacturing to improving agriculture technologies, but also innovate machines with mechanics and smart control develop opto-electronics, sensorics, computer science, and 3D printing. These can also include AI and remotely operated machines like drones or mobile robots that are not yet commercially viable. A combination of these categories of machines leads to independent autonomous alternate agri-robotics. Ultimately, these machines could fully autonomously and steered by AI based software and no need for human supervisor, maintenance and control. These can be simple machines like Mueller robots for weeding.

2. BACKGROUND

This section presents a tutorial on integrating machine learning and data engineering, using time-series predictive maintenance as an example. The tutorial is organized into two parts. The first part discusses the workflow of time-series data engineering, covering data collection, transformation, and storage using cloud-based warehousing, lake houses, and open-source technologies. The second part discusses time-series predictive maintenance models, including pre built and custom models using popular ML and DL frameworks and time-series feature engineering and forecasting techniques. Integrating data engineering and machine learning tools enables end-to-end data processing and predictive analytics pipelines to be implemented in the cloud.

There is a growing interest in integrating machine learning and data engineering as a crucial component of data-driven systems. Successful machine learning projects require sound data engineering to ensure high-quality, timely data for ML algorithms. Integrating ML and data engineering allows continuous monitoring of changing conditions and the retraining of models as new training data becomes available. Additionally, it reduces the total cost of ownership. Most widely used processing engines have home-grown tools that specialize in specific workflows. For example, the DataRobot and H2O.ai cloud platforms democratized machine learning by including the most commonly used package in the box. However, they are not a panacea for end-to-end data pipeline automation. The state of the art can either serve as prebuilt monitoring tools or building blocks for constructing them.

Several case studies and research works discuss deploying forecasting models in cloud platforms or using low-code or no-code tools for training machine learning models. However, typical tools either focus exclusively on predictive maintenance or cover only a sub-topic, such as time-series feature engineering or simple deployment and data engineering tools. To the best of the authors' knowledge, no effort has been made to cover both time-series data engineering and predictive maintenance.

2.1. Overview of Smart Agricultural Machinery

Precision Agriculture (PA) or Smart Farming is a systematic approach utilizing state-of-the-art IoT systems to optimize the efficiency of growing and harvesting crops while conserving resources. Through advanced data collection and analysis techniques, the production of crops is significantly improved. Smart agricultural machinery is at the forefront of this change. These smart systems utilize sensors, cameras, and radars to collect data from the environment, which is processed and analyzed to direct the behavior of farmers and agricultural machinery. Predictive maintenance is a crucial component of smart agricultural machinery and is thus of paramount importance. A farm with a malfunctioning machine can lose vast quantities of crops. It is key to apply the right data engineering and ML models to improve reliability and accuracy.

The significant increase in the social demand for crops has led to the increased adoption of precision agriculture by farmers and experts, and subsequently, millions of intelligent sensors are being utilized to gather data ranging from crop growth to machinery wear state. This results in the generation of Terabytes of data every day, referred to as "Big Data". However, at the moment, this data is mostly not being utilized at all within the farms and companies. In fact, no data science operation is taking place in most farming operations and procedures. Thus, taking on the challenge to develop a smart product - targeting both agricultural companies and individual farmers - to improve their efficiency in terms of crop yield and to assist in preventive maintenance on the machinery is a rather challenging yet promising product.

Smart agricultural machinery is a vehicle consisting of numerous electronic equipment that is designed to operate on fields, such as tractors, harvesters, seeders, etc. The machinery usually features a variety of sensors and controllers and is referred to in its entirety as an embedded system consisting of multiple parts that remotely communicate enabling smart operation with minimal human assistance. In smart agriculture, the machinery senses the conditions of the field to autonomously plan and carry out machine events. To do that, it receives data from the different sensors on the vehicle that are attached to different components of the farm and soil environment. Data is ingested from the opening and closing of hydraulic valves, state of the engine housing, electronic clutch state, etc. Analyses are performed over the data collected from these sources, during which statistics are created on how often each sensor communicates and what is the type of signal produced. The data can then be transformed into useful insights creating a better picture of the conditions of the field and state of wear of the components of the farm machinery.

Equ 1: Feature Aggregation Pipeline.

$$X_t = f(S_t, E_t, M_t, C_t)$$

S_t: Sensor data (e.g., vibration, temperature)

E_t: Environmental data (e.g., humidity, soil type)

M_t: Machinery operational metrics (e.g., usage hours, speed)

C_t: Categorical/contextual inputs (e.g., crop type, season)

f: Feature engineering transformation pipeline (e.g., normalization, rolling average)

2.2. Importance of Predictive Maintenance

Unplanned failure of mission-critical equipment undermines production, and any prolonged downtime results in significant monetary loss. Incremental downtime is thus perennially a growing concern in industries with revenue generating assets such as High-End Servers, Oil Rigs, Wind Turbines or even Nation-Wide Railway Networks. However, unplanned equipment downtime involves not only the risk of loss in potential revenue, but also by-product risks that pose further implications on the performance and safety of other equipment and processes. On the other hand, monitoring and gathering of asset usage and condition data through the strategic deployment of sensors and networks have become an increasingly ubiquitous state of affairs in modern industries. Consequently, vast amounts of collected data from such monitoring systems dubbed the data deluge have become available. Therefore, there arises a critical business need to derive actionable insights from the data for intelligence in mitigating downtime risks and ensuring optimal performance and safe operation of equipment. Such forethought has yielded a new paradigm for Engineering Asset Management termed Predictive Maintenance (PM) originating from Maintenance Engineering. It combines disciplines such as data science, business intelligence, statistics, reliability engineering, and maintenance engineering to manage the uncertainty in event-based maintenance decision making from multiple perspectives. This paper focuses only on equipment-failure based predictive maintenance.

However, acquisition of such a large amount of data poses a slew of fundamental challenges in managing and coordinating the data deluge. Traditional data analytic approaches that rely on domain experts or rules-based conditions become unwieldy as data analytics move from on-dimensional results towards multi-dimensional analysis. Existing machine-learning based predictive models gaining attention due to the promise for timely equipment condition predictions becomes inadequate as it relies on configuring a model architecture and optimal hyperparameters manually. Another dilemma for predictive maintenance applications in practice is the Ask Vs Access divide where sensitive business queries sought by decision makers remain unaddressed and out-of-reach due to the sheer size, scope and unstructured nature of the acquired data. Worst still, a one-size-fit-all approach commonly adopted in many predictive maintenance efforts to be queried in the same manner not only substantially deviates from the business semantics for maintenance decision making but also is likely to overlook important finer-grained queries. Such challenges emphasize the need for understanding why the existing PM efforts failed to generate actionable data driven insights and relevant lessons.

2.3. Current Trends in Machine Learning

The continuous remarkable growth of the ‘internet of things’ has induced device interconnectedness and the generation of unparalleled quantities of data in the last decade. The ‘data economy’ has continually gained traction, as an established winner in the IoT landscape, offering essential opportunities for various industries and business sectors as well as the launch of completely new ones. Businesses are investing heavily in big data collection, storage, and processing frameworks and algorithms, incentivized by the firm’s wide-ranging collection of benefits. In particular, clean, contextualized, and ideally real-time data revelation is defined as a prerequisite prior to the exploits of innumerable opportunities afforded by either ‘data-driven’ analytics to continuously enhance operation and performance or ‘machine/deep learning’-driven predictive insights on machine intelligence. However, the value gained from those systems in harvesting the major investment return is dependent on how compressively and effectively are the data collected, contextualized, mined, represented, and rendered to business insights. While the involved technologies or algorithms are constantly developed and currently form well-established strengthening capabilities, the fabrication of coherent and proficient architecture embracing the aforementioned areas of competence is at its infancy promotion stage, especially at a reasonable operating cost.

Predictive maintenance applications have recently gained increased attention in the IoT domain, where, as a form of condition-based maintenance (CBM), their remark capabilities can contribute to notable business cap rates, return of investment, and competitive advantages. Despite the demonstrated lower costs and chances than conventional failure-based or scheduled maintenance approaches, the adoption rates of modern predictive maintenance schemes are still currently limited in practice. Their viability and cost efficiency have not been established across the manufacturing and service industries, with the monitoring, planning, and mitigation activities being largely conducted offline or mixed with conventional methods. Novel processing frameworks and methods are required to materialize the incorporation of machine learning capabilities directly in the day-to-day operation of the assets so that moderate implementations costs and training times are guaranteed. The huge vocabulary of asset-related elements, modes, and attributes that each system possesses renders the formation of broadly universal processing architectures highly improbable. Therefore, novel iterative resources arrangement and task abstraction techniques are needed to adopt the development of relative processing solutions to the industry’s state-of-the-art.

2.4. Role of Data Engineering

To answer the questions in the introduction, it is essential to gather all relevant information and data, such as failure history (if applicable), logs from machines and devices, configuration files, and the structure of databases and data streams. The following:

- The gathered data can be categorized into types such as time series, flat tables, texts, and images.

- Data quality is examined in terms of absence, completeness, duplication, accuracy, currency, and integrity.

- Automating the estimation of data quality metrics requires checks, filters, and cautionary measures.

- Since data quality is an abstract concept, it is represented with multiple metrics.

- Data quality analysis needs to integrate all available results into a global evaluation score.

In the end, data pipelines retrieve the data and perform quality assessment procedures. This process can be either custom-built or based on ready-made solutions. Since the proposed platform supports PANDA Data Quality, APIs can deliver processed data for data analytics and machine learning tasks if their capability satisfies the data quality requirements.

Data marts are designed to hold data in a denormalized format suitable for bulk operations. They will store streaming data from machine repositories. Qualified streaming data is sent to data warehouses to be aggregated according to user-defined intervals. Qualified data is also pushed to data stores supporting web applications, storing aggregated machine health data in a denormalized format. Thus, reporting and visualization can happen simultaneously with the arrival of the raw data, and machine operators can have access to real-time health data and visualize them in time frames of their choosing. use different methods for data storage. PanDA offers different choices among NoSQL databases and data warehouses for data exchange to satisfy various requirements.

3. MACHINE LEARNING TECHNIQUES

The requirement to innovate and reduce costs aligns with Industry 4.0 trends that prompt manufacturers to adopt new technologies in production processes and supply chains. For this context, predictive maintenance models take advantage of data and machine learning algorithms to overcome such downsides. Several instances detail engineering traditional manufacturing processes to lessen machine downtime. Yet, there is limited research in ancillary agricultural technologies closely tied to climate-neutral proposals. For the EU to become climate-neutral by 2050, farm equipment manufacturers must innovate to enhance the sustainability of their machines and production processes.

Thus, the advent of smart agricultural machinery that is capable of making decisions and applying individual machine learning models to perform those decisions holds potential. However, such innovations incur increased mechanical complexity, leading to greater demands from systems for maintenance. To tackle this matter, new predictive maintenance systems based on machine learning are presented along with an architecture that poses challenges to event engineering and integration of data engineering and artificial intelligence.

In the early 20th century, research found that modern predictive maintenance practices entail risk since machine learning techniques require incorporating the field of data engineering. This has vastly improved the applicability of machine learning models in numerous domains. However, predictive maintenance model integration efforts demand engineering both the complex systems that generate data as well as the machine learning models that infer knowledge from that data. Current predictive maintenance research in agriculture caters to only one side of that challenge, thus a novel approach is proposed.

3.1. Supervised Learning

In supervised learning, the machine learning model is created based on a labeled dataset. Roughly speaking, a label is an answer to a prediction or classification task provided with the training set of the model. The general approach to supervised learning is to treat the previous labeled/answered data points as examples for building the prediction model by replicating the predictive function that was used to generate the labels. Since the true label is not available for working with a prediction model, a commonly used approach is to evaluate its performance with another independent sample. Evaluation metrics include accuracy, precision, and recall, which provide important quantifications of the quality of the predictions and capture different aspects of it. Essentially supervised learning proceeds in two steps: a training one, where a supervised learning algorithm learns a model from a training set consisting of data points with true labels, and a testing/evaluation one, where the model inferred from the training set is applied to an independent test set to assess its performance.

In supervised learning, tasks and applications usually come with different names and descriptions, but the two-step procedure is nearly always the same. The first task of supervised learning is a learning process on data with a supervision signal. Generally, the goal of the learning task is to infer the characteristics of the underlying function mapping from the input instance's space (feature space) to the output instance's space (target space) based on a certain set of observed input-output pairs, also known as examples or training samples. In practice, supervised learning includes a learning mechanism and output decisions mechanism. Accordingly, there are usually two types of output of supervised learning as labels or numerical values forming prior conditions for subsequent data processing.

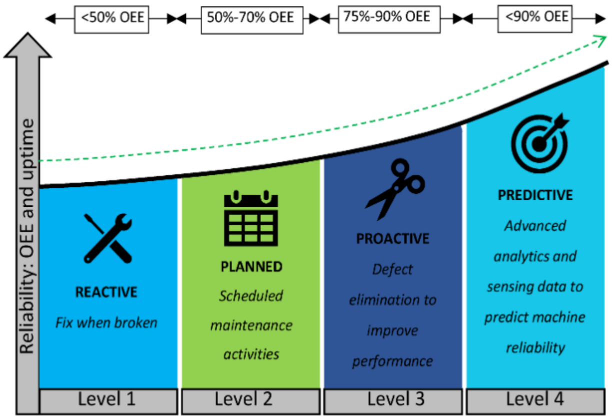
<50% OEE
50%-70% OEE
75%-90% OEE
<90% OEE
Reliability: OEE and uptime
REACTIVE
Fix when broken
PLANNED
Scheduled maintenance activities
PROACTIVE
Defect elimination to improve performance
PREDICTIVE
Advanced analytics and sensing data to predict machine reliability
Level 1
Level 2
Level 3
Level 4

Fig 2: Machine Learning in Predictive Maintenance.

3.2. Unsupervised Learning

The presented study explores the use of unsupervised learning techniques for predictive maintenance. Results show that both approaches can correctly classify between healthy and damaged samples. However, the clustering approach is significantly faster with a preprocessing time of only 10 seconds. A cluster may contain multiple irregularities and only one clustering approach is required to analyze and classify data. In many industries, failures in electromechanical systems can lead to production losses, idle time, and higher maintenance costs. Early fault detection is essential to minimize these issues, and traditional methods are not available for high-dimensional data. Artificial intelligence (AI) techniques are an alternative for pattern recognition on a variety of data sources, including images, videos, and vibration signals. Predictive maintenance is application software development capable of monitoring electromechanical systems, detecting anomalies, and predicting failures using data from low-cost sensors. To assess the performance and predictability of machines, degradation analysis and supervised learning techniques require a large amount of labeled data, which can be expensive, time-consuming, and even unsafe to obtain. As a result, many industries prefer to use unsupervised learning techniques as a natural and effective approach to identify irregularities and assess whether a machine's condition is normal. Moreover, unsupervised learning classifiers are independent from machine health states, allowing them to adapt automatically after disturbances in the environment or system changes.

Autoencoders and clustering are two of the most popular unsupervised learning techniques. An autoencoder is a deep-learning approach based on a neural network that determines a compact representation of the input by forcing the output to be the same as the input signal. It may employ a convolutional architecture instead of a standard feedforward architecture. The compact representation inherits the dimensionality reduction properties of PCA. Moreover, clustering uses unsupervised learning techniques based on a set of n vectors in a domain of d dimensions that cluster independently for high-dimensional data and determines n observed and hidden variables that influence the formation of the ground truth clusters. Moreover, it reconstructs around K clusters at each iteration that merge together to model an unseen sample without re-training the entire model. Thus, clustering can be a natural approach for effective pattern identification after reducing the dimension of the signal data feature space with an autoencoder. Deep and compact clustering are combined to accurately detect irregularities and improve performance and efficiency.

3.3. Reinforcement Learning

When dealing with long-term schedules, an estimated maintenance process must be generated in advance, and the time of maintenance must be reordered and repaired based on an emergency prediction system. The issue of intrusive reordering prevention must also be covered. Most existing Long Term Scheduler solutions do not cover the whole process comprehensively, and each step of the scheduling process, including scheduling, assignment, and transportation tracking, is designed as independent systems. Long Term Scheduler works by computing the best time to carry out maintenance based on both machine and crop conditions from the defined scheduled period. If the situation in the following processes enables urgent maintenance, data is forwarded to the Process Criticizer. The process rerouter changes the maintenance site of down or intruded machines based on machine remaining battery level and distance.

DOI: 10.17148/IJIREEICE.2021.91215

A Reinforcement Learning-based Process Criticizer receives several reports a minute from an Edge-based Processing Segment. Each report is based on either emergency prediction or maintenance completion information. Only reports with a certain degree of credence said by a long-term scheduling agent will be processed, as it is not time-efficient to compare too many reports in the constructive critic system. Based on the trained critic model, the process criticizer will choose one of many reports to re-evaluate. The critical model is periodically updated using regularly collected historical data. Maintenance completion re-evaluations are parameterized as binary classification tasks. Failure case data should be forwarded to the Logical Policy Generator if an urgent maintenance assignment is found.

3.4. Deep Learning Applications

Worldwide, extensive efforts are made to modernize agriculture through the use of advanced information and communication technologies. Precise and efficient high-throughput agriculture requires agriculture 4.0, or smart farming, which enables the continuous monitoring, measuring, and analysis of a myriad of various phenomena and aspects of complex agricultural ecosystems. Smart farming has recently received greater focus, interest and investment mainly due to high labour costs, stringent quality and safety standards, and low margins and efficiency in agri-food chains. The challenges of agricultural production include quantity, quality, nutrition, ecological footprint, environmental impact, food security, and sustainability, which need to be addressed in an integrated and orchestrated way. Agriculture, which is the basis of food production and food supply, is an important economic sector worldwide, comprising a quarter of the worldwide economy. Agriculture is by nature complex, multivariate, and unpredictable, posing a number of risks. Extensive inputs, such as land, water, arable area, fertiliser, pesticides, labour costs, and capital investments are required, posing a number of risks. Each crop and soil type has its own microclimate and pests, with unforeseeable weather events. Nevertheless, agriculture is one of the first sectors which adopted and implemented various technologies starting from the first applications of mechanised machinery. Research institutes and universities invest significant effort in developing, testing and adapting new technologies.

Other hand, the use of computers and the Internet in agriculture, also referred to as e-agriculture, can significantly enhance the existing tasks of management and decision/policy making through context/situation/location awareness and might assist in better understanding the rather unpredictable agricultural ecosystems. In order to enhance management in agriculture, ICT needs to be used to continuously and accurately monitor, assess, manage, forecast, and remotely control various assets and processes in crops and farming. Correspondingly, the multitude of tasks in agriculture is data- and knowledge-intensive, which require appropriate models, standards, methods, algorithms and technologies.

4. DATA ENGINEERING FRAMEWORK

In the domain of predictive maintenance, the lack of standard data representation format or protocols has been reported to be one of the roadblocks for a more wide adoption of prognosis capabilities in multiple industries. In this regard, a framework offers a wide variety of services and a standardized data model. The proposed solution, preventive maintenance architecture in flexible manufacturing using the framework, addresses issues in predictive maintenance analytics. The architecture solution is proposed first, followed by prototype implementation and experiment results. After that, development, testing and endpoint details of a co-simulation prototype are described. After describing the implementation, runtime performance analysis and evaluation results are then presented.

The framework provides various services for building smart applications, such as context data management; stream and complex event processing; analytics and graph; etc. These services obsensurate desktop applications for different domains. The proposed solution aims to provide a flexible and modular architecture using the framework. The flexibility is achieved by facilitating the integration of different components, each required for predictive maintenance. The proposed 4-layer approach supports better understanding of different components/processes at different levels as well as the overall architecture. Interoperability of different components as pluggable components can be easily integrated with the designed predictive maintenance solution. This feature will enable effective maintenance analytics with minimal effort. The adopted context broker along with the Big Data Analysis module facilitates interaction and integration of existing as well as future IoT devices within the production plant. The application of connector along with the proposed data model will facilitate greater interoperability and transparency of data access. It will facilitate virtual factory production mode, requiring a higher level of data integration from customers, suppliers and partners across enterprises, optimizing the information flow and delivery process.

Equ 2: Predictive Maintenance Model.

$$\hat{Y}_t = \mathcal{M}(X_t;\theta)$$

$\hat{Y}_t$: Model output (e.g., failure probability or time-to-failure)

X_t: Input feature vector at time t

θ: Model parameters (e.g., from a random forest, XGBoost, or LSTM)

4.1. Data Collection Methods

Studies on predictive maintenance show that evaluating the usability for various industrial machines requires an extensive analysis of underlying data. Such exploration requires the extensive understanding of sensors, the data they generate and the means of processing that data. Furthermore, available datasets have to be analyzed in detail and the analytical tools have to be understood thoroughly. Since a number of IoT datasets from real-world usage of various industrial machines are publicly available, they are used for this purpose.

To ensure that the selected machinery satisfies the operating conditions and allows for prediction properly, a detailed understanding of various mechanical components and their operating specifications is required. Moreover, finding the right level of details in which the prediction should be done is not trivial. The level of detail refers to what type of failure is of concern; is it failure in a bearing, in a gearbox, in a controller, in a motor or in any other driveline equipment? On the contrary, the level of details refers to the straight prediction of the time until complete failure in the whole machine. Automated diagnostic and prediction systems would therefore require time and performance data to determine the type of condition monitoring equipment.

Time-series data from industrial machines has to be carefully considered, especially when estimating the remaining useful lifetime of engines or predicting the failures in car sensors. Continuous operation of machinery is a requirement and designated standards on the failure event count has to be followed. Moreover, the time-series data needs to be properly published, meaning that the sampling rate, amount of missing values and noise level has to be reasonable and understood before usage. Reinforcement and supervised learning as prediction engines can later be trained and labelled with predictions, which requires a direct mapping from threshold deviations to predicted failures.

4.2. Data Processing and Cleaning

The following analysis describes the data-related efforts completed to develop the ML solution for Dozer, one of the Use Cases in the PELAGOS project. The data consisted of six years of operation of the Dozer machine in several fields in Northern Finland. The dataset is recorded by a higher-end telematics unit that gathers a wide multitude of telematics parameters, all recorded every 5 min (in some cases every 1 min if a parameter changes status). A preliminary client-driven filtering of the data also took place, with records containing non-working hours and irrelevant parameters dropped (an average of 311 TelOG records per day per telematics parameter).

For the ML efforts, the dataset has to be further trimmed (either manually or automatically) to remove anything that doesn't meaningfully connect with the preventive maintenance theme. The key local indicators (KLIs) from the telogs were identified to make clustering possible (in practice, a lot of effectively similar parameters were dropped). This resulted in about 71 parameters written in some proprietary code. Some of these include the working hours of the arm, fuel consumption, fuel changing, and working hours of engine resources, among many others.

Unexpected gaps and outliers in the data were almost ubiquitous and had to be handled. Machine learning (ML) methods often require extensive preprocessing and cleaning of data to eliminate errors and inconsistencies. As non-lab-based processes, the datasets gathered from many sensors typically include problems such as erroneous measurements, incorrect data types, repetition in the records, odd-value signals, or many consecutive NaNs (missing values). The cleaning efforts were therefore structured into three parts: data analysis-driven cleaning of the outlier values, selection-driven data-cutting based on shape; and value-driven NaN imputations.

4.3. Data Storage Solutions

The main goals of the data storage solution are to effectively support the envisioned data workflow compared to alternative prior works, and to serve as a prototype architecture for future research endeavors. In addition, supplementary goals consist of providing postgraduate students an opportunity to participate in related research for career building, to demonstrate the accessibility of data reinventions in smart agriculture with hosted open-sourced components, and to develop and document a build procedure for the architecture that can be communicated to farm tech companies. These goals are addressed through the proposed architecture, which is flexible, modular, and extensible. The core components of the predictive maintenance application are held within FIWARE's NGSI-LD and IoT Agents, Elixir-based Chronix,

DOI: 10.17148/IJIREEICE.2021.91215

and Grafana, and integrated into a functional architecture using docker compose. The architecture is hosted on a digital ocean droplet to ensure reliability and accessibility. The components were selected for extensibility, documenting standardized queryable endpoints with open-source codebases for use in extending the application. The EasyExe and Docker EXE data ingesters are originally built to allow customization for specific data storages, while the refurbishments to the other components are contained within FGAD- and fixed GO-API- repositories. An extensive access guide is retained in the repository README for adapting or building the architecture and working with the components individually. A video demonstration summarizes the development process and shows off the implemented components. Given the rapid pace of technological advancement and agricultural development, it is anticipated that the current implementation will need renovations in future research. Sufficient documentation and an online hosted architecture aim to minimize the friction of revisiting and revising the data storage solution.

4.4. Data Pipeline Architecture

The ETL process performed by a scheduled container is defined in the following steps. The implementation of this architecture follows the NGSI-LD standard for web APIs, ontology model definition, and data interchange. The scheduling of ETL tasks is implemented through the combination of Crontab and a python script. The data integration steps are implemented based on the Python libraries and available NGSI-LD API libraries.

Before the ETL steps are described, the architecture is operated in a cloud environment. Data publishing in the NGSI-LD format is used to connect this architecture via HTTP(S). Multiple users in different locations can access the architecture via their web browser without restrictions of the operating system platform. The whole architecture may need at most twenty minutes to restart after any failure belongs to server facilities, which is feasible for users with new releases every five minutes. Cybersecurity protection measures are employed to protect against common network-based attacks on computers and networks.

An ETL task is defined as follows: 1. Connect to the Orchestrator platform Redis database using the python API to access all structured data. 2. For each uniform resource identifier (URI) of the machine data model, 3. Generate new acquisition based on a model and data at the current time including the available, set points, and operational conditions values. 4. Store all new data in the Orchestrator context broker using the NGSI-LD format. 5. For each machine, one or more messages are stored in the Message Queues of the Orchestrator forwarded to each monitoring dashboard. 6. Close the Redis connection and complete the ETL process.

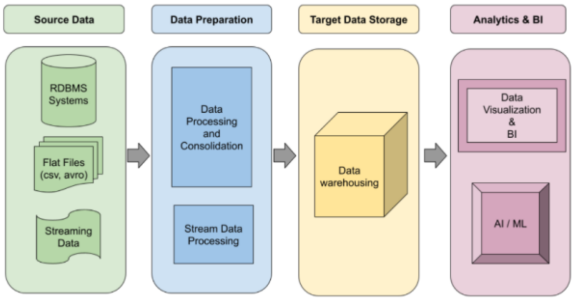

Fig 3: Data Pipeline Architecture.

5. INTEGRATION STRATEGIES

Even though the casual nature of analyses needs extra knowledge like domain-related engineering or manufacturing knowledge, it can be adjusted for a better performance while properly introducing prediction maintenance techniques that could adapt for those other domain applications. These techniques rely on how the data acquisition and measuring missions have been carried out. Two main scenarios can be considered, the first one related to data acquisition systems and their constraints, and the second operational data, as they could be of value regarding system global maintenance policy but at a different positioning and analysis level. Scenario 1: Complete data acquisition systems A strategy regards maintenance methods aimed at anticipating system failures when taking into account remaining useful lives instead of only reacting to failures when they occurred at costs of delays and high repairs; so, education of predictive models based on the respective dataset and KPI indicators is of utmost importance. Initiatives generally combine different methods, models or techniques, regarding the complexity and the number of monitored KPIs and signals. The solution is non exhaustive as needed tools add accuracy but makes touching it and using them harder. Relying on the adaptability of traditional data processing tools to trade-off between performance and simplicity of use by non-specialised operators is essential. Therefore, few tools have, up until now, been directly implemented and included within intelligent maintenance systems such as it was developed within the machining centre. It is a built-in solution that maintains, monitors, interfaces real-time data with operators while adjusting to any specific CNC type and configuration. It is based on a smart cloud-based data acquisition system and engines running independently of the tools as sort of services. This tool is fully operational on this machine range and is mature to be extended by including new tools such as much more focused on one specific KPI.

5.1. Combining Machine Learning with Data Engineering

Despite the benefits of digital transformation, many heavy machinery manufacturers worry that they may miss out on the next technological frontier if they develop an incorrectly calibrated ecosystem or build one that no longer fits the requirements of their business. Well-established companies, which subsequently form families of brands providing equipment, increasingly face the challenge of utilizing data from existing machine ecosystems and harmonizing it with data from future machinery for delivering the promised smart applications. There is currently no comprehensive vision on how to efficiently combine data generated by machines operating on a heterogeneous, brand-diverse, and technologically outdated landscape. With the ecosystem still in flux and supporting standards being formed, it is still unclear what baseline solutions can be built today, and how they need to adjust in the future, should data-based applications mature. The absence of a unified language across various brands and the heterogeneity in technical implementation have rendered these questions even more challenging. The aim is to clarify the baseline capabilities of smart maintenance for tractor and machine brands evolving in isolation through heterogeneous data engineering, a universally applicable machine learning value chain for deriving and executing knowledge from raw data, and a frame for assessing the results. The frame is anticipated to serve as a basis for informing subsequent developments of the smart maintenance approach by specifying the information needs and reflecting on the requirements for an application. It can also be utilized for comparative benchmarking of solutions and capabilities. Prior to describing the frame and its application, the results are presented, specifically the analysis of the processing of existing data. Focusing on the investigation of actively generated short- and long-term raw data for understanding machine health. Translating existing knowledge on machine health and creation of indicators to this machine domain and examination of how maintenance knowledge is created from the status indicators. The general machine learning value chain is illustrated and translated to the machinery maintenance domain. An initial version of the frame for evaluating existing smart maintenance capabilities at machinery brands is mutually derived and applied for a case study on two grass harvesting machines. The analysis of results is presented and used as an entry point into the problems facing brands. Last but not least, future research needs are outlined, like the visualized learning aspect of the frame and a large-scale application of the updated formulation.

5.2. Real-time Data Processing

Using a message broker, it's possible to develop a real-time data processing application to simulate the Smart Agricultural Machine system's working environment. For example, become a publisher that makes a random number generator in another application, and publish the data generated with random integer values from 0 to 1000 at intervals of 1 second for the ROP server and the task. This data can be generated as the topic according to the necessary document. In other applications, subscribe to this topic using the same broker, and display the published data on activities to ensure that data is successfully exchanged between applications. The takeaway messages of the data processing module are receiving real-time data through the MQTT protocol to prepare data for predicting the condition of Smart Agricultural Equipment. Predicting results are also published on the message broker to use for display maintenance history.

Using ROP and SCADA topics as a setup data storing application, ROP and SCADA servers subscribe to the message broker. When the Real-time Operative Data (ROP) is received, it is recorded in the CSV file by the ROP server at the same time. Afterward, the data that is recorded for 2 hours is sent to the maintenance history display application as CSV files and saved in the monitoring application's directory file. The rest of the generated ROP data is saved in the same data_csv folder as the ROP server, and this data is sent to the predictive maintenance algorithm. Besides receiving real-time data from MQTT protocol, SCADA server also subscribes to Task topic as a demanding operational task to do. It saves the received data to the local SQL database, and displays comprehensible statistics of that data, such as dark red color indicates to do an immediate task.

5.3. Scalability and Performance Considerations

The data engineering component of the IoT pipeline uses Apache Kafka to buffer high-frequency sensor data to Amazon Web Services (AWS) Redshift. The machine learning model is deployed with the FastAPI web framework asynchronously. Given the separation of data treatment, model retraining, and node health reporting, this component is easily restated and modified. The algorithm is designed as a two-component ensemble model to handle the trade-off between prediction accuracy and performance. Each predictor uses an independent model class, which streamlines and

DOI: 10.17148/IJIREEICE.2021.91215

facilitates the addition of alternative models and hyperparameter exploration in the future. Consequently, it is easy to adapt the V-Arima component or change its loss function to compute the MAE or RMSE.

Both predictors use adaptive feature extraction methods and are pre trained with more than half of the training time periods to reach the performance target. The simple model with less time-consuming feature engineering is expected to be faster than an XGBoost or MLP based prediction model once the one-shot feature selection is performed. The performance of pretraining should be tested more carefully with other seasonalities, thresholds, or hyperparameters. Nevertheless, most datasets benefit from the configuration, and the pretraining performance is within acceptable boundaries when not satisfiable.

In addition to the level-wise evaluation, different aspects of modelling choices need to be treated further, leading to auxiliary questions summarized in Appendix B. Besides the predictive model, other modelling choices to be determined include adjustments for periods with several skipped executions, dimensions for temporal or feature splitting, configuration of the feature engineering library, and node characteristics to analyse. Most of these choices span all modelling cases, but the chain of repeated characteristics provides the opportunity to semi-automate time-consuming parameter tuning.

6. CASE STUDIES

In this section, case studies will be presented to uncover the potential of machine learning and data engineering methods in prediction analysis tasks using data acquired from real operating agricultural machinery. The case studies involve prediction analysis tasks from daily soil tillage and soil making operations, including soil tillage ridging prediction, vibration feature extraction, prediction of the vibration of soil tillage machine components, and prediction of micro-stoppage behavior of a soil making machine. Recent years of research work on these tasks and their resulting solutions will be explained and discussed. Various architectures of machine learning solutions developed in-depth, and they are used to illustrate the integration of AI and data engineering solutions.

Soil Tillage Ridging Prediction Analysis Task One case study is to predict the formation of ridge soil tillage type operation soil spillage over time. A soil tillage machine employed to pull-under soil tillage operations which randomly forms ridge type shape operations is used to extract operating machinery signals like implement working speed, vertical movement of the implement, preliminary operating signals, vibration signals of the tractor and hydraulic oil cooler, and accelerometer signals on the implement. Application relevant data extraction simulation and uncertainty quantification methods are developed and employed to address the signal reliability challenge. Then time-series advantages of novel data engineering methods for timeline association, cropping/re-arranging, and handling data loss are proposed to bridge data engineering with machine learning methods flexible to work with the learning features extracted from both time domain and frequency domain. Time-series based prediction analysis solutions are proposed as this is a task to predict an event happening when a rising region is formed on an associated time-series data. These proposed solutions accurately predict the ridging event using its constitutive feature as input without using domain knowledge. The integration of data engineering methods bridge design and implementation of widely used LSTMs as plug-and-play learning blocks. This case study uncovers the potential of interpretable prediction modelling to raise critical feature inputs with statistical significance on the predicted event.

Vibration Feature Extraction Analysis Task A high-vibration situation of soil tillage machinery is difficult to predict in advance and prevent due to the inconsistent cause of high-vibration phenomenon. This case study is to detect the high-vibration situation of hydraulic shovels on a soil tillage bin which is a critical high-vibration part of the soil tillage machine. A soil tillage machine employed to pull soil making operations is used to extract vibration signals, operating machine parameters like working speed and error codes, and environmental factors lifted from the ground working friction power and humidity. A spatially informed multi-channel CNN is designed based on an insight gained by the analysis of soil making and relevant stochastic processes sampling data and enabling the captured behaviour for long-term prediction analysis and interpretation. This solution transparently detects the machinery's high-vibration behaviour with high accuracy and manages convolved filters to perform local representative frequency band search. This case study uncovers the value of explainable and transparent machine learning results in the potential application of real-time indicator on-season soil making operations and future machine collaboration between soil tillage machinery and data pipelines.

Equ 3: Maintenance Decision Rule.

$$\delta_t = \begin{cases} 1 & \text{if } \hat{Y}_t \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

6.1. Predictive Maintenance in Tractors

Predictive maintenance identifies the condition of in-service equipment and determines the maintenance actions required for a given period. This aims to minimize maintenance costs while maintaining a high level of safety and machine availability. Predictive maintenance is widely adopted in industries for complex machinery and critical components to avoid failures and delays. The predictive maintenance system consists of data acquisition, data pre-processing, predictive maintenance model generation, predictive maintenance execution, and failure diagnosis.

Predictive maintenance is implemented in several SISO machines and process automation. A twin-screw extruder is analyzed using regression and classification models to estimate the remaining useful life of the machine using temperature, pressure, and motor current measurements. A critical transmitter used for pressure and temperature measurement is monitored for six different types of sensor failure. In an electro-mechanical system, predictive maintenance is adopted to model the current's temporal effect on the predictive maintenance decision. The modeling of predictive maintenance using event logs is discussed.

Several approaches use output from sensors recording physical phenomena linked to the degradation process of the machine. In some applications, event logs generated by the machine are used instead. The predictive maintenance problem is formulated using the event logs of a fleet of automatic teller machines generated over several years. A new publicly available ATM event log dataset is introduced. As a case study, the CAPA method is applied to this dataset to perform predictive maintenance on an ATM fleet. A different approach to solving the predictive maintenance problem is also introduced, illustrating another use case on the same dataset. Finally, other interesting avenues related to the use of event logs in predictive maintenance are highlighted.

6.2. Sensor Data Analysis in Harvesters

The analysis of the data collected from sugarcane harvester sensors is performed using machine learning (ML) techniques. A set of seven input variables is defined, constructed from the data that can be measured in the usual operational situation of the harvester. Six algorithms are trained and compared, including Linear Regression (LR), Artificial Neural Network Multilayer Perceptron Backpropagation (ANN), Support Vector Regression (SVR), K-Nearest Neighbor (KNN), Extreme Learning Machine (ELM), and Random Forest (RF). The implementation of a fluidity or mass flow measurement in sugarcane harvesters has considerable economic potential for companies, farmers, and sugar and ethanol plants. To obtain advantages from mass flow data, it needs to be merged with a Model Predictive Controller (MPC) capable of managing the company's system’s leveling and gauging control. Data monitoring involves the storage, organization, and annotation of a requested data set over a certain period. The data examined through analysis is formed mainly as time series measurements. Due to the amount of data generated and the speed of their processing, designing the software architecture to observe the sensor signals is a challenge. In this software architecture, the processing and continuous storage of production data need to be designed to be carried out off-site when they generate a greater amount of data than what can be managed online. The analysis results based on ML techniques are expected to allow for the detection of drift or anomalies and subsequent processing. The output of the data analysis consists of information for monitoring and alarm decision-making, with time references being essential. However, the diversity, format, and frequency of collected data present challenges to integrating this information into a coherent view of the monitored environments.

6.3. Irrigation Systems Monitoring

The intelligent monitoring of irrigation systems is a critically significant task for agriculture development and the national economy. The soil moisture of a reasonable constant level is crucial to guarantee the proper growth of crops, and attaining proper soil moisture depends on the design and control of irrigation systems. In general, soil moisture is controlled by executing certain irrigation behaviours, which is a challenging task, especially when the irrigation systems are large-scale and complex. A typical large-scale irrigation system consists of watershed areas with complex variations in geography and climate conditions, and at the same time, it involves a wide array of behaviours, including irrigation design, systems design, and systems control.

DOI: 10.17148/IJIREEICE.2021.91215

Solar-powered intelligent irrigation machines mounted in a vehicle-based form factor that can work on outdoor terrain for a field-scale study is provided. Semi-automated processes with minor human interventions such as preparation of operation mode parameters and observation of machine movements may prevent failures in control. A demonstration of site-specific irrigation with remotely observed flood irrigation across a conventional furrow-flooding basin irrigation system is provided. Initial testing and further information on the machine operational system are included in this demonstration. A methodology to guide the machine designs and estimate operating parameters of the field-scale smart irrigation machines is also proposed. The proposed operational mechanisms and techniques can be adapted as self-sufficient and low-cost solution options for many similar problems that agriculture and farming face today. The currently available commercial solutions tend to be complex and expensive.

Having limited scalability across regular and degraded states of operational environments, such limitation creates barriers to the adaptation of advanced control machines to many conventional agriculture systems and practices.

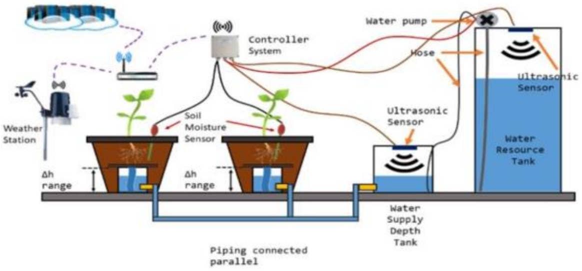

Fig 4: Irrigation Systems Monitoring.

7. CHALLENGES AND SOLUTIONS

Predictive maintenance (PdM) has gained attention in a variety of sectors, because it enables enterprises to enhance equipment uptime, reduce costs, and optimize overall productivity. Many researchers have worked on PdM from various perspectives since then. Data-driven models and machine learning (ML) techniques have produced highly accurate models when enough quality training data are available. ML adoption in oil and gas operations is hindered primarily by data, culture, and strategy issues. Data-related issues such as availability and quality of data, auditing of data, feature engineering by domain experts, interpretability of algorithm outputs, and security issues are challenges for the adoption of AI/ML technologies.

IoT applications generate more data than can be used in conventional data processing systems, leading to a storage crisis as well as the need for increased processing power. Stream processing architectures overcome this challenge through the design of modularized data processing topologies that consolidate data queries and actions into a real-time processing workflow. However, the development of stream processing topologies is hard for end-users. The difficulty handling the velocity of streaming data limits predictive and defer processing in situations where the model cannot run in-realtime due to excessive delay associated with the training loads or inference. Finding an integration method that would enable data engineers to deploy ML models in a streaming architecture without needing specialization knowledge is also a challenge.

The development of stream processing architectures has facilitated the management of data generated at data rates and volumes higher than can be reasonably stored. Although these applications are becoming more common at the same data velocity, the integration of pre-trained global ML models with Kafka remains unexplored. An end-user oriented integration approach makes it easier to deploy ML in-stream data processing topologies.

7.1. Data Quality Issues

Sensor data communications and connections are critical in MIL-ICSs. Wire inside the tunnels need to be protected from harsh environment, corrosion, wear and tear. Exposed TCP wired connections are susceptible to interference and hacking due to the opening of the ports, including TCP and UDP. Keeping the security of MIL-ICSs is of utmost importance to not just the operational life of the military platform but also the safety of equipment, soldiers, and the covert nature of military operations. Sensor failures deteriorate the situation awareness of outside environments, vehicle driving state, weapon engagement information of cyber attack, etc. Understanding, predicting and diagnosing the cause of these failures can prevent severe degradation of the mission effectiveness of the military systems.

DOI: 10.17148/IJIREEICE.2021.91215

Data from MIL-ICSs can be categorized into two types: offline data and online real-time sensor-streamed data. The offline data includes persistent data in text, CSV files and databases. The structured offline data is easier to format than the semi-structured sensor status-data. Various data pre-processing methods such as handling missing values, outlier removal, normalization and discretization are available for structured data cleaning. The online sensor-streamed data are in json format. The challenge is on how to properly parse the incoming stream data to get all needed information. The difficulty arises from messy sensor streams, which include redundant and dirty information. Because of entrance to the parsing stage being in the format of streaming data, some data parsing and cleaning techniques such as dropping of semi-structured data is not applicable here.

The MLiS approach extracts novel physical information from streaming sensor data superseding the traditional counter between two sampling points, which is named and considered as the first derivative of sensor data. Second derivative soon before a sampling point, named and considered as a jerk, is also extracted. These two compound features depict the force and acceleration of vehicles from knowledge of physics, being able to accurately and effectively detect the pushing and tilting of the infrared target. However, jerk extraction is an irreversible transformation. A careful consideration has to be made to determine how to handle this sensor data one-shot at parsing stage.

The need for eliminating unexpected tool breakage, unscheduled downtimes, enhanced product quality, and more competitive business effectiveness, transforms traditional run-based maintenance to zero/dynamic time maintenance, which is an overarching customer-demanded option. From the data edge, recent aero-sensor developments, ever-cheaper sensor deployment options, and the rising convergent data availability from SCM, CAD, Materials Handling Systems, ERP, MRP and CLM, can be utilized for zero-downtime, and in conjunction with proper logical models, to achieve immediate feedback of the detected incursions to immediate planning or even back to the design stage.

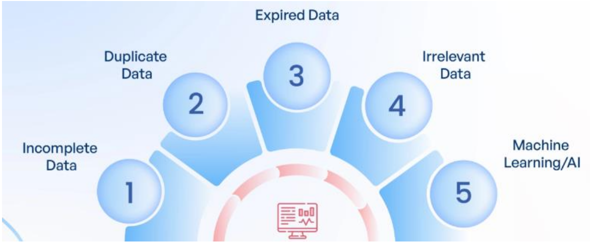

Fig Data Quality Issues

7.2. Model Accuracy and Reliability

Predictive maintenance (PM) is steadily becoming a mainstream technology across numerous industries and domains. Given the recent explosion of data in many process industries, data-driven maintenance and, specifically, machine learning (ML)-driven PM have attracted great interest from both academia and industry. However, this interest has not yet been reflected in wide adoption. This paper aims to investigate the current status of research and development of the cross-field PM, focusing on the three core sub-domains: data strategies, modeling strategies, and explaining strategies. Numerous promising solutions are presented as well as challenges, gaps, and future directions concerning each sub-domain. Finally, landmark papers are surveyed to highlight essential benchmarks and trends in each sub-domain.

Artificial Intelligence (AI) refers to the ability of a software or hardware to imitate intelligent human behavior. With no clear boundaries, AI includes many facets such as deep learning (DL), machine learning (ML), knowledge graph, and natural language processing among many others. One of the incoming areas of disciplines under the umbrella of AI, is predictive maintenance (PM).

7.3. Integration Complexity

The proposed system integrates multiple architectures and platforms. The implications of the integration of every single platform are detailed in this section. Regarding the Mobile Web Application, the main focus of the development was on creating a centralized and intuitive dash. Indeed, a web application that retrieved and showcased the information about the performance of machine equipment, scheduled maintenance activities, tasks for the maintenance operators, and dashboard settings was developed. The web application was designed to be fast-loading, targeting to reach complete satisfaction of educational and non-business users. Additionally, performance versus security is a delicate trade-off area where many business applications are vulnerable. The development team executed a substantial amount of testing during the entire development of the web application. These tests were both performed: operating system tests, internet speed tests, error handling tests, etc. Furthermore, the accessibility of the domain, the dashboard URL, and the monitoring URL was tested.

The Integration of the FIWARE case on the proposed platform was mostly focused on generating a receivers hook that consumes the “Event” data sent by Perception. Precisely, a microservice was developed that receives events through a webhook, catalogs them in a MongoDB database, and sends an alert to the Pusher API and the dashboard. Its development was not overly complex; the most demanding parts were those involving Pusher and web socket management. It should be noted that, due to lack of time and API documentation, some FIWARE API calls were not fully implemented. The integration of the fleet management microservice consisted of a case router microservice applied to manage communication among Differential Location Detection APIs and the Process ID Broadcaster API. In addition, a fleet management microservice was developed to keep track of machine equipment activity and interaction. It channeled state information from an activity event through a Pusher channel.

8. FUTURE DIRECTIONS

In the realm of predictive maintenance (PdM), machine learning (ML) applications are evolving rapidly, a quiet revolution inspiring new research ideas and newer PdM mechanisms . At first, early implementations were bolted on top of ad-hoc preventive maintenance schedules. As research expanded, new broadly defined PdM architectures emerged paying homage to long-proven data-driven modelling methodologies. The research blossomed into diverse techniques addressing complex PdM challenges but rapidly creating a jungle of obscure definitions, methodologies and tools to fully explore the impact of ML on PdM.

Smart agricultural machinery has created large volumes of data worsening the ability to translate data into actionable information. Making machinery and farmers smart provides an opportunity to create a new pool of agricultural data on machines incorporating advanced sensors and data engineering technologies adding real-time intelligence to transform agricultural machinery. ML-powered approaches, along with data engineering processes, can optimize machinery performance and bring predictive maintenance capabilities to provide accurate state and system health monitoring that can affect productivity. Towards that goal, the integration of ML and data engineering for predicting maintenance is advisable paired with social and economic models to study their impact on agriculture’s operation and benefits. Future predicted data engineering patterns such as continuous event archiving in delayed cycles and time plateaus where structural covariate modelling of the data changes could be ground-breaking. The aforementioned ML workflows and tools can handle the temporal data and find usable patterns for automatically knowing the “what,” “how,” and “why” of machinery failures.

Collaboration with domain knowledge engineers can bring even more advanced methodology with multiple knowledge bases, definition sets, visualizations, and rule determination models bridging the gap between the digital and physical worlds.

8.1. Advancements in Machine Learning

In recent years, the integration of machine learning and the internet of things has gained significant popularity in data science. In this thesis, the use of machine learning models combined with internet of things devices is examined in the context of predictive maintenance for agricultural machinery. Here, the motivation and context behind the research questions are presented. Agricultural machinery, like all types of machinery, undergoes wear and tear over time. Failure prediction is referred to as the prediction of wear and tear or machinery failure. Such predictions can be used to perform maintenance on the machinery before failure occurs. Since bad failure predictions can easily affect the financial stability of a company that maintains machinery, even small optimization in failure predictions can be of great help.

To better understand the need for and possible applications of predictive maintenance of agricultural machinery, some trends in agriculture and the agricultural machinery market are presented. Significant technological advancement has been made in agriculture over the past few decades. Increasingly, more types of agricultural machinery have been developed. As a result of the technological advancements and increased complexity, predictions of wear and tear within agricultural machinery are needed. This will assist technicians, farmers, and manufacturers in taking preventative actions.

The predictive maintenance pipeline for agricultural machinery is divided into three parts. First, data engineering needs to be applied to obtain meaningful features. Data engineering, in this context, is the gathering, storing, and cleaning of data, as well as transforming it into readable formats. Second, a machine learning model is selected and trained. The trained models will then be selected using a set of evaluation metrics. Third, the selected models are subject to time-based predictions to see how long before failure predictions take place. In addition, the selected models are compared to a benchmark model.

ISSN (O) 2321-2004, ISSN (P) 2321-5526

IJIREEICE

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 9, Issue 12, December 2021

DOI: 10.17148/IJIREEICE.2021.91215

The predictive maintenance pipeline is then utilized to make predictions on gathered data from agricultural machinery. Multiple models are trained to make predictions based solely on signals collected from sensors from machinery. The models tend to take into account the information of a certain time window style. Thus the state-of-the-art log-based model cannot be directly integrated into the current pipeline, so only the supervised classification models are trained using the data from agricultural machinery. Over a dozen different variants of these models are trained, but eventually, only two models are selected and compared.

8.2. Emerging Data Engineering Technologies

The emerging data engineering technologies are described in the following subsections, including "Cloud Computing" for data storage, "Edge Computing" for real-time data processing, and "The Internet of Things" for inexpensive sensors. As for the efficient machine learning technologies, "Federated Learning" for model training using data at the edge, and "Algorithm-Data" based methods for smarter and more efficient machine learning algorithms, are introduced.

They are orthogonal to one another in a sense that they can enhance or augment one another, for example, edge computing is suitable for federated learning and the Internet of Things can provide more data for anomaly detection.

Cloud computing is a data-intensive technology which provides on-demand network access to a pool of configurable computing resources. As a mainstream computing infrastructure, it is well known in two forms: platform service (PaaS) and stored all data in the cloud. For the predictive maintenance task of the smart agriculture machinery, an open-source cloud computing engine named FIWARE with both PaaS and SaaS as APIs can be adapted and extended. The collected data streams from many data sources such as sensors in the smart agriculture machinery are stored in a Fiware powered cloud database, called "Orion Context Broker". On the other hand, the cloud virtual machine is provided as a service for developing machine learning models for time series forecasting and writing results in the cloud too.

Edge computing is also a distributed computing paradigm that brings computation and data storage closer to the sources. It is capable of pre-processing images in the mobile devices, real-time data processing and lower data latency (more than 70% reduced latency). Equipped with various sensors for monitoring environment and vibration of equipment, the edge devices can gather a wealth of data. As a primary computing/deployment environment, edge computing is best for real-time anomaly detection. Since model training is cumbersome and resource demanding, a pre-trained machine learning model for detecting abnormal signals can be transmitted to the edge devices for anomaly detection. The results will be sent back to the cloud side together with raw signals when an anomaly is detected, so that a further analysis can be carried out in the cloud.

8.3. Potential Impact on Agriculture

Precision farming and data-driven (DD) agriculture are arguably providing the future for the agri-food supply chains with the enough resilience to answer both: the ever-expanding planet population and the continuous changes of the environment. Data is seen as the main driver to turn this vision into reality, with geospatial and non-geospatial data at its heart. This is the case for data-centric agricultural research as well: the global vision is built on consolidated statistical and multi-sensor platforms that allow for persistent monitoring of both: the crop and the environment, unprecedented informatics techniques, but also pixel-wise, time-dense geodata exploitation.

Sensor data represents farmers' first-level input to this data-driven revolution. It is becoming increasingly available with commodity prices, research, and agri-tech company satellite and aerial imagery providers. As a result, the data streams are expected to always grow, even more with future satellites carrying new environmental sensors. Predicting and avoiding these events has far-reaching consequences for agriculture, environment, food security, and precision farming. The agricultural sector is largely dependent on weather, soil conditions, and evaluates multiple conditions to arrive at a decision. Rain-fed crop irrigation and perfect disease-pest management remain challenges.

On top of these premises, the current availability of high-performance computing (HPC) infrastructures on cloud is again democratizing access to advanced modeling methods and big data. At the same time, with new data collection and storage technologies, it is often cost-effective to generate and collect data instead of implementing new modeling efforts. Nevertheless, there are still ambiguities in the data-driven epiphany: majority of the accuracy of the models, which can be achieved only with investment in the data. Data is coming from heterogeneous sources and various modalities, giving rise to the so-called information overload, a no trivial barrier under productive use of agri-data and induced practical and theoretical challenges alongside great opportunities.

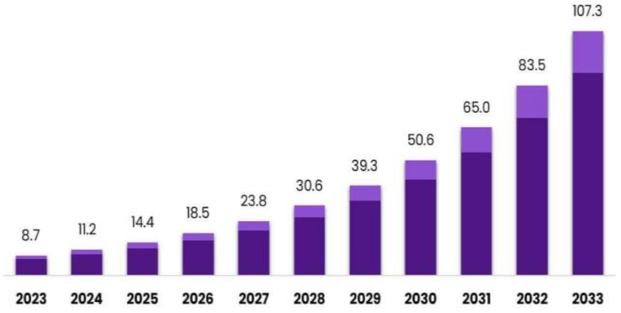

Fig 6: Integrating Machine Learning and Data Engineering for Predictive Maintenance in Smart Agricultural Machinery.

9. CONCLUSION

This paper presented a methodology for integrating machine learning and data engineering for predictive maintenance for smart agricultural machinery using a case study of a precision planter. It was shown how data engineering techniques as well as insights gained from domain knowledge and computer engineering processes are important for preparing data for model training and how these processes can significantly enhance the potential of machine learning algorithms. A detailed overview of the applied data engineering techniques was presented.

The proposed approach was evaluated on the case study of a precision planter where machine learning and data engineering methods were integrated to develop a prediction model of faults in a hydraulic system. Failure data was collected from ag-tech companies and pre-processed using various data engineering techniques. Several machine learning techniques were applied and benchmarked, showing the method can significantly reduce maintenance time and allow for a more efficient workload allocation between technician resources. It was noted that through the presented approach, significant insights were gained from the domain knowledge and industrial context. Apart from the applied machine learning model, the gathered data engineering techniques can be valuable for further case studies and algorithm benchmarkings.

The research aimed to enhance the predictive maintenance for smart agricultural machinery, for which a real-world case study of a precision seed planter was used as a representative example. The employed methodology for data acquisition and engineering as well as the applied machine learning algorithms for fault prediction were presented. The research shed light on the critical domain knowledge of smart agricultural machinery and demonstrated the necessary data pre-processing steps and model selection and evaluation details. This research is expected to foster future research in the area while being immediately implemented in practice.

REFERENCES

[1] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29–41. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1294

[2] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55–72. Retrieved from https://www.scipublications.com/journal/index.php/ojms/article/view/1295

[3] Avinash Pamisetty. (2021). A comparative study of cloud platforms for scalable infrastructure in food distribution supply chains. Journal of International Crisis and Risk Communication Research , 68–86. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2980

[4] Anil Lokesh Gadi. (2021). The Future of Automotive Mobility: Integrating Cloud-Based Connected Services for Sustainable and Autonomous Transportation. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 179–187. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11557

[5] Balaji Adusupalli. (2021). Multi-Agent Advisory Networks: Redefining Insurance Consulting with Collaborative Agentic AI Systems. Journal of International Crisis and Risk Communication Research , 45–67. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2969

[6] Singireddy, J., Dodda, A., Burugulla, J. K. R., Paleti, S., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Universal Journal of Finance and Economics, 1(1), 123–143. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1298

[7] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101–122. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1297

[8] Gadi, A. L., Kannan, S., Nandan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87–100. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1296

[9] Cloud Native Architecture for Scalable Fintech Applications with Real Time Payments. (2021). International Journal of Engineering and Computer Science, 10(12), 25501-25515. https://doi.org/10.18535/ijecs.v10i12.4654

[10] Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967

[11] Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures.

[12] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. American Journal of Computing and Engineering, 4(2), 35-51.

[13] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. Available at SSRN 5102662.

[14] Chinta, P. C. R., & Karaka, L. M.(2020). AGENTIC AI AND REINFORCEMENT LEARNING: TOWARDS MORE AUTONOMOUS AND ADAPTIVE AI SYSTEMS.