International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Integrating Machine Learning and Databricks for Work Order Automation in Pharmaceutical Manufacturing

Name: Srikanth Reddy Katta

Email: skatta304@gmail.com

Abstract: The operation of manufacturing pharmaceuticals and drugs requires high efficiency, accuracy and definitive compliance with standards set by regulatory bodies. Work order management, which is a critical and generic area addressing how manufacturing tasks are planned, assigned and executed, is still a difficult area given its dependence on both human and automated factor inputs. This paper aims to look at the implementation of ML algorithms and Databricks within SCM and work order management of a pharmaceutical manufacturing firm. This research achieves better task scheduling, anomaly detection, and predictive maintenance by using Databricks' distributed computing features and ML models. The resulting work describes a case study applying both supervised and unsupervised learning to historical data, process modeling and bottleneck detection, and static and dynamic scheduling. The experimental study employing real-world datasets suggests a specific 30% times saving regarding task completion and a general 20% augmentation of compliance with regulations. This paper thus discusses the impact ML-driven automation can bring to the pharmaceutical Industry while considering issues like data privacy and model interpretability.

Keywords: Work Order Automation, Pharmaceutical Manufacturing, Machine Learning, Databricks, Predictive Maintenance, Data Analytics.

1. Introduction

Pharmaceutical manufacturing involves the application of severe legal requirements and guidelines for complicated and delicate operations. Timely management of the work order has an important role to play in smooth functioning, quality assurance and compliance with established regulations. It is evident that most manual work order management leads to time wastage, mistakes, and sometimes even delays. [1-4] Automation has one of the most significant impacts on the manufacturing process. It can minimize human intervention, thus minimize errors, improve productivity and meet the laid down GMP guidelines. Automating work orders is a process of scheduling tasks, controlling them, and even supervising the implementation process alongside handling them. Thus, modern methods of Machine Learning (ML) and advanced data analytics show potential for solving these problems.

1.1. Benefits of Work Order Management

- **Access to Request Details:** The workflow of work order management can also be beneficial in that it provides a common setting where teams can efficiently access the details concerning specific tasks, such as requirements, necessary materials, and timelines, as well as directions. This ensures that all the details needed by the workers and managers when handling tasks are easily seen, and hence, the right work will be done. By reducing uncertainty and imprecision in the communication of intent, the teams can get cracking with work, avoid making decisions based on incomplete or unclear information, and work faster than they have to spend their time looking for vital information.

- **Organized Processes:** This is because a well-structured work order system enforces best practices across the organization and provides a workflow to which teams have to adhere. In it, the complicated job gets divided into simpler tasks; with the help of role segregation and responsibility and even time division, it makes sure that every task goes through a certain pattern. Such simplification of working processes not only contributes to the correct organization of tasks but also to avoiding duplicative work, miscalculations, and the free distribution of resources. This, in turn, results in increased efficiency or output, fewer mistakes, and more cohesion within an organization.

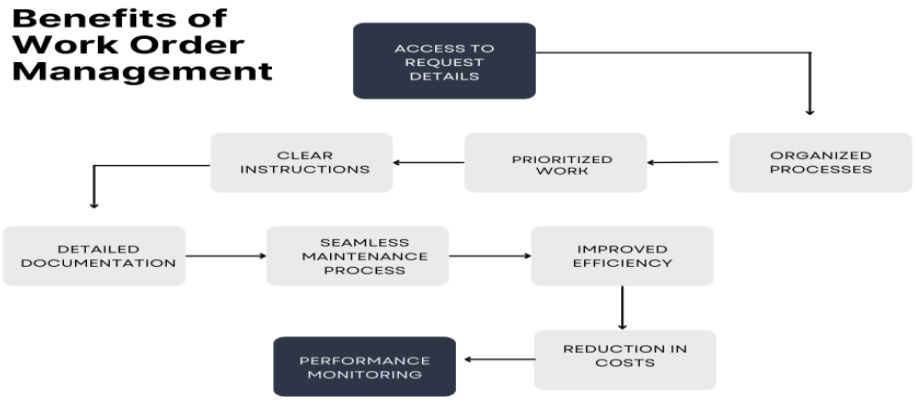

Figure 1: Benefits of Work Order Management

- **Prioritized Work**: Using work order management, people in a team can easily determine priorities such as urgency and importance, how the work is to be done, and which resources are available for deployment. Then, it is possible to sort out the tasks by importance; for example, when the car needs urgent maintenance, this job will be given more time and attention than a less urgent customer support call. This kind of prioritization helps meet deadlines, make critical equipment available for use, and enhance the overall utilization of the available resources, hence enhancing productivity for the customers.
- **Clear Instructions:** Most of the work orders also contain elaborate instructions and procedural requirements that enable the workers to appreciate the depth and width of their work, the steps to be followed, and the standards expected of them in regard to quality and time. This helps to reduce any chances of misunderstandings, mistakes, or lateness that may be occasioned by difficulty in following directions or lack of clear instructions. The ability to refer to assigned tasks' details allows employees to make work decisions independently and with confidence, which saves time and ensures compliance with specific guidelines. In turn, clarity contributes towards getting consistent results across different groups of employees.
- **Detailed Documentation:** Work order management systems also involve documentation, making it easy for an organization to document important aspects of a particular job, including the description of the order, materials used, time consumed, and persons involved. This detailed record-keeping gives an organization an accountable paper trail that increases adherence to organizational guidelines or rules and/or external rules and regulations. Also, comprehensive documentation makes reporting easier, as well as budgeting and planning for the future, as it allows for equal measures in efficiency tracking and enhancement.
- **Seamless Maintenance Process:** Work order management is the maintenance planning and organization that makes every aspect, from preventive and corrective maintenance to booking and performance, easy

and efficient. Through the management of maintenance schedules, the organization is capable of preventing probable cases before they occur, lessening incidences of unexpected equipment breakdowns, and increasing the useful life of the assets. It also helps prevent disruption and makes a big difference in increasing organizational effectiveness.

- **Improved Efficiency:** Work order management systems eliminate manual and monotonous work as well as eliminate data entry errors. In other words, by coordinating issues like the process of allocating tasks and deadlines and managing available resources, more can be achieved in the same amount of time. More observation of job progress also equips managers with the ability to correct mistakes or undue holdups in work progress. This leads to early task completion, reduced operational disruptions and, in essence, increased efficiency.
- **Reduction in Costs:** Work order management systems have been instrumental in cutting expenses by optimizing the usage of resources, reducing time lost and avoiding unnecessary expenditures. Scheduled servicing helps avoid expensive repairs, whereas continuous, streamlined tracking of operating conditions helps avoid wasteful employment of labor and material resources. Further, the frequency of activities associated with task management means less overtime and less waste, which lowers expenses. These combined advantages improve the financial status of the organization and drive up the ROI.
- **Inventory Control:** Among the key success factors in any enterprise, management of inventories is paramount, and work order systems are very useful since they provide real-time information on the status of materials, tools and equipment. The enhanced documentation of stocks helps prevent shortages, which slow down the work process or lead to overstocking, which results in wasted resources. This real-time visibility means resources can be better planned and managed, procurement costs are lower, and the correct material is on hand when it is required, making operations more efficient.

1.2 Role of Databricks in Data Analytics

Databricks is a business that has adopted big data analytical solutions through the processing of big data at scale, speed and efficiency. [5-7] Being a unified data analytics platform based on Apache Spark, Databricks expands data engineering, data science and machine learning functionalities with improved performances, collaboration and scalability. Here are the broad areas in which the firm Databricks fits in the data analytics industry.

- **Unified Analytics Platform:** Databricks is a single unified analytics platform that brings together data engineering, data science, and machine learning. This integration streamlines all stages in the data analytics life cycle, and organizations can take data from their initial ingestion phase right through to the analytics phase and model deployment. Databricks effectively centralize such functions, and this makes communication possible. The elimination of barriers between teams increases productivity. This eliminates the issues of dealing with several tools and technologies, reducing the sophistication of data analysis and making it easier for data engineers, business analysts, data analysts, and data scientists to work with.
- **Scalability and Performance with Apache Spark**: Apache Spark is at the heart of Databricks and is a fast and scalable open-source data processing engine. Spark enables Databricks to process petabytes of data in real time. This is important for industries that operate on big data, such as finance, healthcare, and manufacturing. Exploiting in-memory computations through Apache Spark, Databricks re-creates the analytic pipeline and provides a possibility to perform preparatory steps, including data transformations and aggregations, much faster than traditional tools. This scalability helps the business to be able to handle large amounts of data as the business continues to expand, thus enabling it to compete with other large businesses as data continues to increase exponentially.
- **Collaborative Environment for Data Teams**: A particularly valuable component of Databricks is that it is designed as a collaborative workspace where data teams can collectively build projects. The notebooks for collaborative analysis, version control, and connectors with other tools such as GitHub and Jupyter

make Databricks collaborative for data scientists, engineers, and analysts. It provides the capability to share results, proof documents, and build machine learning models in one environment and in real-time with other teams. It does this by allowing for more collaboration between teams, as well as making it easier to maintain quality across teams and to push out frequent updates or creative solutions.

- **Data Integration and Ingestion**: With regards to data integration and ingestion, Databricks provides a simple solution. With it, you are able to query file data stored in different locations such as Amazon S3, Azure Data Lake, and Google Cloud Storage, among others, and file data in various formats such as CSV, Parquet, Delta Lake and more. Databricks connect with various sources, including databases, IoT sensors, and streaming platforms, and are enhanced by built-in connectors. It also brings flexibility to business, where data pipelines can be centralized and where more generalized data is gathered from multiple sources, businesses can make better decisions.
- **Delta Lake for Data Quality and Reliability**: Databricks also has an Integral open-source storage system named Delta Lake, built on Apache Spark, that improves data lake data availability and quality. ACID (Atomicity, Consistency, Isolation, Durability) transaction support means Delta Lake's data operations are guaranteed to be sturdy. Others are time travel or the ability to query data from the past and schema enforcement, which acts as a backup measure to prevent mistakes brought about by a changing schema or naughty data. Since Delta Lake guarantees optimum storage for data to be analyzed, analyzed data will be of great quality, and this creates confidence with the column industries such as finance and health, which originate their decisions from the data analysis carried out.
- **Advanced Analytics and Machine Learning**: Databricks offers a complete environment for sophisticated analysis and machine learning. It includes interfaces that can integrate with familiar industry-grade machine learning frameworks like TensorFlow, PyTorch, and Scikit-Learn; data scientists can create and train them within the environment. Also built-in, Databricks offers tools for ML, such as MLflow and talks about the ML life cycle, from the experimentation to the model deployment process. Combining machine learning with data engineering, Databricks streamlines the onboard process for employing those predictive models as a way to capitalize on data for predefined uses like predictive upkeep, sales predictions, and outlier detection.
- **Real-Time Data Processing:** Due to its being designed to process data in real-time, Databricks is a beneficial platform for applications that need fresh insights. Using streaming analytics, real-time dashboards, or event-driven workflows, Databricks allows companies to capture data streams and gain insights as soon as possible. For instance, manufacturing industries can use real data collected from IoT sensors to analyze the performance of machines and make predictions of their failures. With data processed and analyzed in real-time, Databricks enables organizations to act quickly to adapt to the environment's condition without delay.
- **Cost Efficiency and Optimization**: As such, Databricks provides an optimal method of addressing the costs linked to working with big data by offering clients scalabilities and manageability over resource utilization. As an operating model, organizations can adjust resources from Small to Large or vice versa through GCP while incurring the cost of the use of the compute and storage only. This characteristic of elasticity is especially advantageous for those organizations that have varying workloads in data processing. Automatic tuning tools are also provided for further performance fine-tuning of the data pipelines with respect to excessive computations or resources. Therefore, scalable and cost-optimum solutions are at the core of Databricks' approach, which allows organizations to control big data analytics costs without overspending.
- **Secure and Compliant Data Management**: For example, healthcare, finance, and pharmaceutical industries always consider data security and compliance with the law very important. A significant factor of security is addressed by Databricks due to the fact that it provides role based access control for the user,

data encryption and support for cloud security standards all through the stages of data processing. In addition, using Databricks brings flexibility to meet many different regulations, including GDPR and HIPAA, and therefore, it can work with various industries with particular strict demands. This focus on security and compliance also allows businesses to rely on Databricks for their most sensitive data, all things considered.

- **Democratizing Data and Analytics**: Databricks provides the tools that enable organizations to achieve data democracy so that more people can leverage complex data processing and machine learning essentials. This means that data engineers, analysts, scientists, and even students are all capable of entering a query, manipulating data, and creating machine learning models. This democratization of the data tools makes it possible for the decision-makers to be in a position to use data analytics in arriving at decisions, irrespective of the lack of data science skills needed to work it out. In this way, Databricks democratize the means of acquiring and applying advanced analytics throughout the organization.

2. Literature Survey

2.1. Work Order Management in Pharmaceutical Manufacturing

The management of work orders in the pharmaceutical manufacturing process has historically been through Enterprise Resource Planning (ERP). These systems have been essential in adding efficiency to processes, defining resource distribution, and enhancing schedules. [8-13] Still as acknowledged, ERP systems have fundamental restrictions in the kind of information processing they can effectively support, to cognizance of real-time information flows and to adapt to the dynamic evolution of production processes in the related setting. These systems are normally legacy-oriented; that is, they depend heavily on past performance as well as fixed decision facts, factors or rules and hence cannot easily adapt or prepare for an exception situation in real time. Thus, using ERP systems enhances the general management of the processes; however, it does not offer what is needed in terms of flexibility the system's ability to foresee or at least identify broken processes early before they affect the manufacture of pharmaceutical products.

2.2. Machine Learning Applications in Manufacturing

Some of the following areas of application reflect the use of ML in the manufacturing industry: Predictive maintenance, Quality control and Process optimization. Studies show how the application of ML algorithms can be useful in minimizing equipment downtimes and increasing efficiency in the event of a failure or breakdown by predicting the same in advance and increasing the quality of the final product. Thus, ML models can take historical and real-time data and, together with that, patterns of failure and patterns of the process. The process itself can be optimized at a level significantly higher than traditional methods. Nevertheless, although the proposed methodology has produced favorable results in the identified tasks, there is still considerable room for employing ML for work order automation in pharmaceutical manufacturing. Currently, there is no particular emphasis on using ML for some essential aspects of work orders management, scheduling, and monitoring of operating real-time processes, which, if improved, can drastically transform both productivity and compliance as essential aspects towards achieving optimal performance within the context of pharmaceuticals manufacture.

2.3. Databricks in Big Data Analytics

Databricks based on the Spark engine are known for their ability to process and manage large datasets as well as for effectively accelerating machine learning operations. The company Databricks has discussed how it has been applied in many fields like finance and healthcare, which require real-time data processing and handling and processing a huge amount of data. Due to its capability of performing large data computations at once, big data analytics is well-suited for Hadoop. In the global pharmaceutical manufacturing sector, the opportunity to use Databricks is quite large to enhance work order automation, which involves a large volume of operational and

production data. However, the author of this paper has not fully researched it in this particular industry. Here, Databricks shows great promise for the pharma industry, driving efficiency in data pipeline processes, model training times, and real-time decision-making, especially within compliance purview.

2.4. Gaps in Existing Research

Despite the tremendous advancement in the use of machine learning and big data analysis across different industries, the implementation of these technologies for the automation of work orders in the manufacturing of pharmaceuticals has not been well explored. Current systems do not integrate well with the large amount of generated data and do not address the flexibility of manufacturing processes. Furthermore, many traditional systems lack the right level of dynamism and flexibility to be compatible with compliance systems that are important when dealing with drugs in the pharmaceutical industry. The existing studies mainly study single pieces of manufacturing optimization, which cut across the problems in the current research; they do not provide a comprehensive framework that combines anomaly detection, real-time decision-making, and regulatory compliance in work order management. This is great finding the gap and, at the same time, underscores the importance of more research and development for enhancing the use of ML and big data to fully automate the management of work orders for the pharmaceutical manufacturing industry.

Methodology

3.1 System Architecture

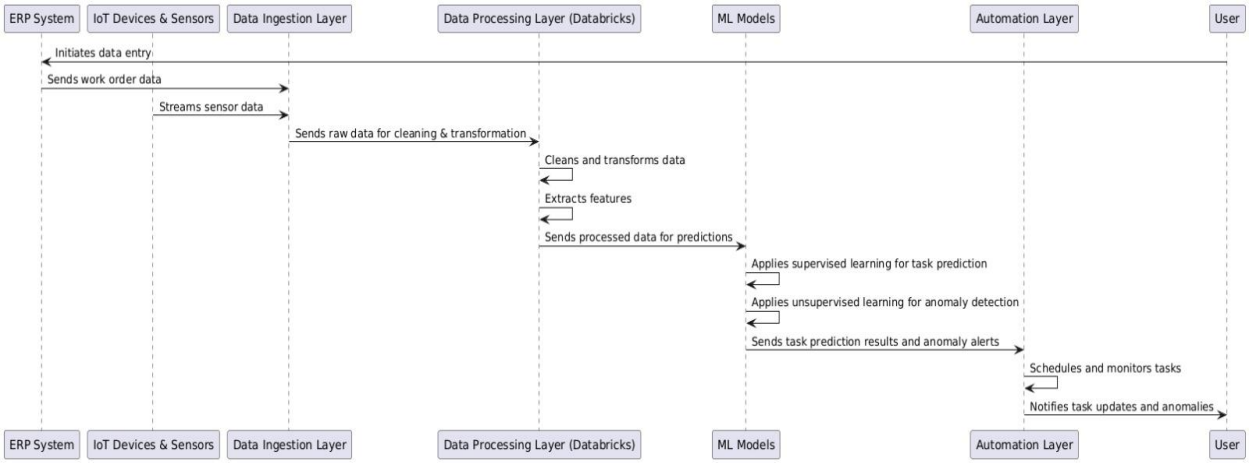

Figure 2: System Architecture

- **ERP System:** The ERP System is used as the starting point and controlling mechanism for the actual operating processes in an enterprise. [14-17] It enables data input and produces work orders, which are, in effect, initiators of other operations. These work orders include structured business data, such as the details of the task, time, or resource details, which are essential. As the middle tier of an integrated ERP System, it guarantees the creation, tracking, and transfer of all required operational data to other systems, closing the gap between business operations and other technical processes. Through it, it minimizes inefficiencies and errors associated with these traditional methods.
- **IoT Devices & Sensors**: Things in IoT, such as devices and sensors, extract real-time information from the physical world. They track the characteristics of the machines, the environment, or the process and constantly report the values of the monitored sensors to other systems. The data includes temperature, pressure, vibration levels, and equipment health indicators, to name but a few. The continuous availability

of this data guarantees that systems contain appropriate and current information to evaluate working efficiency. IoT devices serve as the internet backbones through which real-world data is collected by organizations to monitor and optimize their processes.

- **Data Ingestion Layer:** It acts as the first receiver to collect all raw data coming from various sources, including but not limited to ERP systems and IoTs, for consolidation into the Data Ingestion Layer. Its major function, therefore, is to manage a healthy and efficient consumption of large-scale data that movement does not transfer data un manifesto. The information generated from work orders in ERP and from sensors of the IoT are collected and preprocessed for further cleansing and normalization. Consequently, the Ingestion Layer serves as a buffer and passes this data to the Data Processing Layer, where such composed data is indeed performed. It affords a significant contribution in relation to instigating and Che-Shape-secting a healthy knowledge-tiding structure for data flow.
- **Data Processing Layer:** The Data Processing Layer, executed often on big data platforms such as Databricks, is the layer where the data is processed through an important sequence of operations to obtain actionable insights. This process involves preprocessing, which consists of both data cleansing and conversion to fit the use of the analysis. It also involves removing all aspects, such as inconsistency, errors, and noise. This is succeeded by feature engineering, where numerical and tangible characteristics are obtained from the data for input to machine learning algorithms. In other words, although it may bear some characteristics of a ‘dirty’ data source, it is clean enough to be usefully fed into various types of downstream predictive and analytical tasks. Owing to the large and complex structures of today’s large datasets, platforms such as Databricks, with its scalable and cloud base capabilities, make it easy for businesses to process large datasets in a short span of time with increased accuracy and draw timely information that when acted upon can make a big difference.
- **Machine Learning (ML) Models:** Machine Learning (ML) Models are a key component that underlies predictive analytics and anomaly detection, and they have to use the processed data to provide solutions. These models employ two primary approaches: Supervised Learning, as the tasks of forecasting timelines, recognizing inefficiencies in the workflow, or predicting activities based on data history, and Unsupervised Learning, as the identification of irregular patterns, trends, or behavior, including machine performance issues or operational declines. Based on the analysis of data and the detection of valuable patterns, the ML models generate and restore the results of task prediction jobs and send out anomaly alarms that are important for real-time decision-making and optimization. These outputs are automatically transmitted to the Automation Layer for further performance of actions, improvement of operations, and resolution of problems proactively.
- **Automation Layer:** The Automation Layer uses information and results arising from machine learning models to enhance or recommend a process. It autonomously assignates and oversees works relative to the task prediction result, helping to assign the most appropriate work flows for the given tasks to be accomplished effectively. Also, it sings out for anomalies and alerts organizations to respond to such challenges before they emerge. The automation layer also provides a way of interfacing with users through alert messages to the users on the progress of a task or on the identification of an unusual occurrence. Such interaction in real time minimizes the use of human involvement, increases efficiency and puts organizations in a position to address issues unique to an organisation's operational aspects.

3.2. Data Preprocessing

- **Handling Missing Values Using Mean Imputation**: In this step called data preprocessing, missing values are responsible for data gaps and can be dealt with in four ways. Mean imputation is the most frequent kind of imputation in which the subsequent missing situations in a numerical column are replaced with the mean value of the column in question. This approach preserved the relative amount of spread in the data and the

discarded records that are generally incomplete. Proper handling of missing data means imputation retains a solid and strong data set ready for use in subsequent analysis.

- **Normalizing Data Using Min-Max Scaling**: This data normalisation technique normally transforms all the result numerical values into a fixed resultant range [0, 1]. This method scales each observation by subtracting the smallest value in the feature and then dividing it by the range, which is the difference between the largest and the smallest value in the feature. Another advantage of normalization by Min-Max Scaling is that it does away with the problem of the differing scales or units, which gives the algorithm an easy way to process the data. This step is rather critical, especially for feature-dependent models such as neural networks or distance-based algorithms.
- **Extracting Features Such as Task Duration, Machine Status, and Operator Performance**: Feature engineering is the process of searching for simple and complex aspects of large databases and turning them into more usable features for learning models. It consists of features that give information about operational efficiency and time, the duration of a task, the condition, and the performance of a machine during a task via the machine status feature. Second, efficiency and effectiveness measures, like a completion rate or an error ratio, can be used to assess the operator's input in the processes. By defining these features, much more informative data is extracted to be used by models in making accurate predictions, identifying patterns and improving processes to the best optimum.

3.3. Machine Learning Models

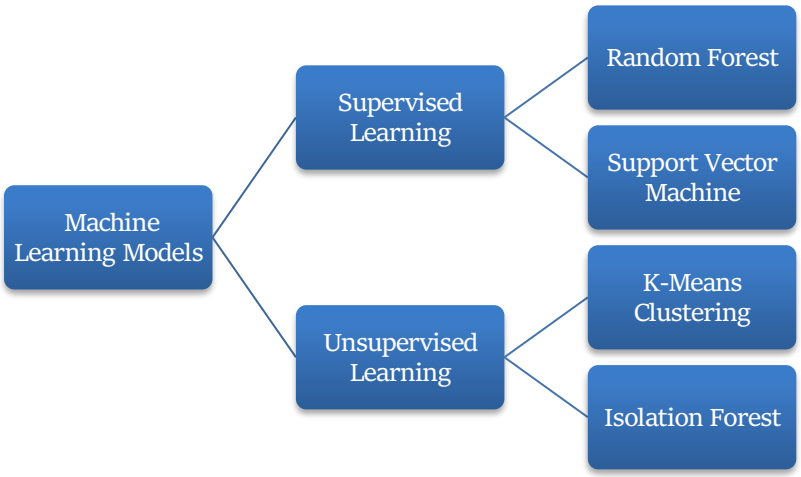

Figure 3: Machine Learning Models

3.3.1. Supervised Learning

- **Random Forest:** Random Forest is one of the families of supervised learning algorithms typically used for regression problems like task time estimation. It constructs many decision trees on some portions of the data and then synthesizes the results from those trees to produce precise predictions. Random Forest also overcomes the problem of overfitting by aggregating results from multiple trees; thus, it is ideal for large, noisy data. [18-20] In this context, it helps predict time, which is taken to accomplish the activities, where from historical data, an organization identifies the patterns to schedule the work adequately.
- **Support Vector Machine:** SVM is a type of supervised learning algorithm used dominantly in the classification process. It does it by determining the best-fit decision surface (or hyperplane in the case of a linear classifier) between these data classes. Furthermore, when it comes to the priority of the work order, SVM involves features of the tasks like due date, type of work contained in the order, and the available

resources in order to categorize the orders according to the priority as either high priority, medium priority or low priority. It assists in the organization of matters related to the operational running of the business in a way that makes it easy to identify and solve problems that affect it.

3.3.2. Unsupervised Learning

- **K-Means Clustering:** K-Means clustering is a clustering algorithm used for partitioning data sets into clusters based on accomplishments of concepts known as features. It defines the values of the actual data to the corresponding cluster centers and repeatedly realigns the boundaries of clusters to best minimize the variance within the clusters. In the context of tasks, K-Means clusters tasks by organizing some characteristics like the task duration, resources needed, or type of task. This grouping helps organizations establish work similarities to identify patterns that can be used to reduce embellished processes or identify resource patterns to meet organizational needs.
- **Isolation Forest:** The Isolation Forest is also an unsupervised learning algorithm that is clearly invented for anomaly detection. It does this by partitioning the data randomly and then comparing how many partitions would be needed to separate an example into different classes, where less would be required for an anomaly. They revealed that, out of 28 clustering algorithms, Isolation Forest is quite computationally efficient and highly effective at outlier detection. In the case of task management or monitoring machines, it recognizes deviations from normal work, patterns of delay, abnormal behaviors of machines or irregular usage of resources, all critical factors that an organization can correct early on.

3.4. Implementation of Databricks

Databricks Delta Lake, MLlib, and MLFlow create a strong and highly scalable ecosystem to work with big data and train and deploy ML models. Delta Lake, developed on Apache Spark, is used for reliable data ingestion from cloud storage, databases, and streaming platforms with guaranteed data consistency, quality, and immutability under atomicity, consistency, isolation, and durability (ACID) transactions. It enables both micro and batch data processing at once, as well as time travel queries on the data and schema evolution for use in a change in the data structure. Made for Apache Spark, Databricks' MLlib can work with Big Data to deliver classification, regression, clustering, and much more. It supports distributed training in the decentralized approach and hyper-parameter optimization for the actual effective models.

Last of all, MLFlow, an open-source platform hosted in Databricks, helps track and manage all the phases of the ML model from the initial testing to the production. It is used for managing experiments, versioning, and putting them into production in the form of REST endpoints for online/ batch prediction. The Model Registry in MLFlow will allow only the best models to be deployed for production use and provide monitoring, retraining, and governance. All these tools help to support the whole machine learning flow, beginning from data intake and model development to deployment and management, allowing for working at scale.

3.5. Evaluation Metrics

- **Mean Absolute Error (MAE):** In the case of regression-based models, the metric that is used is the Mean Absolute Error (MAE), which is the mean value of the absolute differences between actual and predicted values. It is calculated using the formula:

$$\text{MAE} = \frac{\Sigma \mid \text{Actual} - \text{Predicted} \mid}{\text{N}}$$

N stands for the number of data points from which the data will be taken. The MAE is one of the easy-to-calculate performance metrics for a regression model that defines how close the actual values are to the predicted values.

Since it takes the absolute differences, large errors are not punished the same way as with a squared error, which can be beneficial to determining the average size of the model's error.

- **Precision:** Precision in the context of classification based models is commonly used to estimate the certainty of the positive predictions being made by the model. It is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Of each of these categories, the correct positive instances are True Positives, while the incorrectly predicted positives are False Positives. The best suited is when the consequences of false positives are costly, and Precision optimizes the positives, thereby reducing the likelihood of overpredicting the positives.

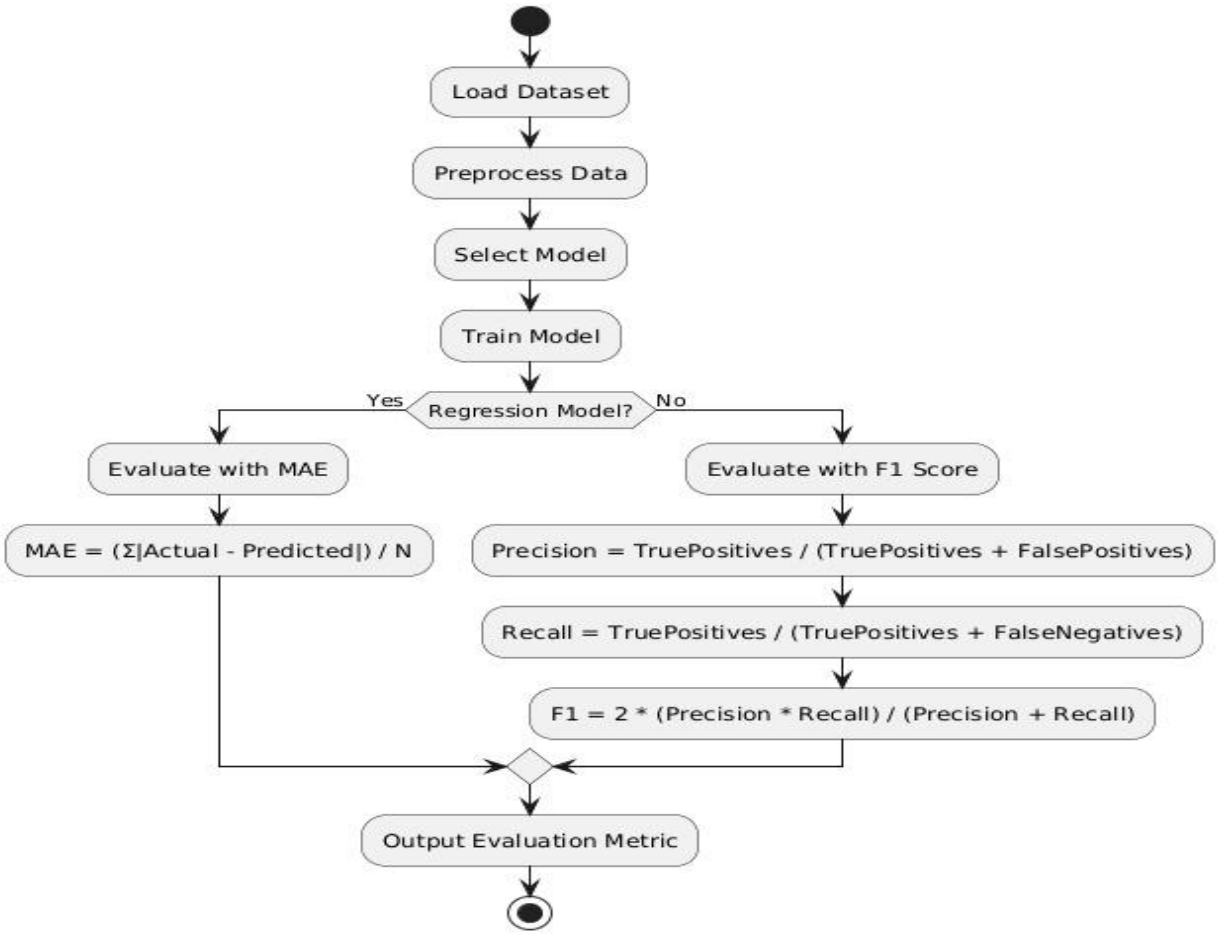

Figure 4: Evaluation Metrics

- **Recall:** True positive or sensitivity also measures the passport ability of the model to correctly identify all the positive cases. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Here, False Negatives mean what the model forgets to predict as positive actually are positive cases. Recall is important when false negatives or missed positive outcomes are expensive, for example, in screening a disease, so that as many true positives as possible can be yielded.

- **F1 Score:** The F1 Score is simply the harmonic mean of the Precision and Recall working as a single metric when the classes are unbalanced or when the Precision vs Recall trade-off is important. The formula for F1 Score is:

$$\text{F1 Score} = 2 \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}$$

F1 Score is the average of Precision and Recall with an emphasis on the smaller value. This makes it especially optimal, for example, in cases where not only the type I and II errors need to be managed but also a broad perspective of the classification performance is required.

4. Results and Discussion

4.1. Performance Metrics

The efficiency and effectiveness tests of the proposed system used MAE and F1 Score, which showed an increment of effective scores after the application of the machine learning techniques.

- **MAE (Mean Absolute Error):** Mean Absolute Error (MAE) is one of the essential measures used to measure the performance of a system in terms of its predictability, especially when the predictability estimates time durations or completion estimates. In this context, the MAE in the traditional system was 15 minutes, indicating that the time estimates from the system were, on average, 15 minutes detached from the actual time taken to complete the task. However, the proposed system that incorporates a machine learning approach reached a value of MAE of 8 minutes. This reduction by 7 minutes only shows that there is an addition and accuracy in time predictions as a central business tool that would contribute to more order and correct rationing of resources where a business is actively in operation. An improvement of the order of 36.93% of the baseline MAE not only establishes the accuracy of the proposed system but also points towards the reliability and consistencies of more accurate and less varying task time estimates, which is of paramount importance in dynamic working scenarios.
- **F1 Score:** The F1 Score is another measure commonly used to confirm the model's efficiency, in the course of the training process, especially if there is a major difference between the measures of Precision and recall. It gives a resource that harmonically gives the accuracy in being positive the result while accurately giving the amount and nature of instances (recall). Thus, the F1 score for the traditional system equaled 0.75, which indicates a fairly good ability to navigate between Precision and recall. The performance of the proposed system indeed improved significantly, with an F1 score of the verbs set to 0.88. This enhancement suggests that the new system is better equipped to identify and classify tasks or anomalies as accurately as possible in terms of minimizing false positives and false negatives. This way, the obtained F1 score increase demonstrates its generalization capability as well as increased reliability for the tasks that are more complex and time-varying.

Table 1: Performance Metrics Comparison

Metric	Traditional System	Proposed System
MAE (minutes)	15	8
F1 Score	0.75	0.88

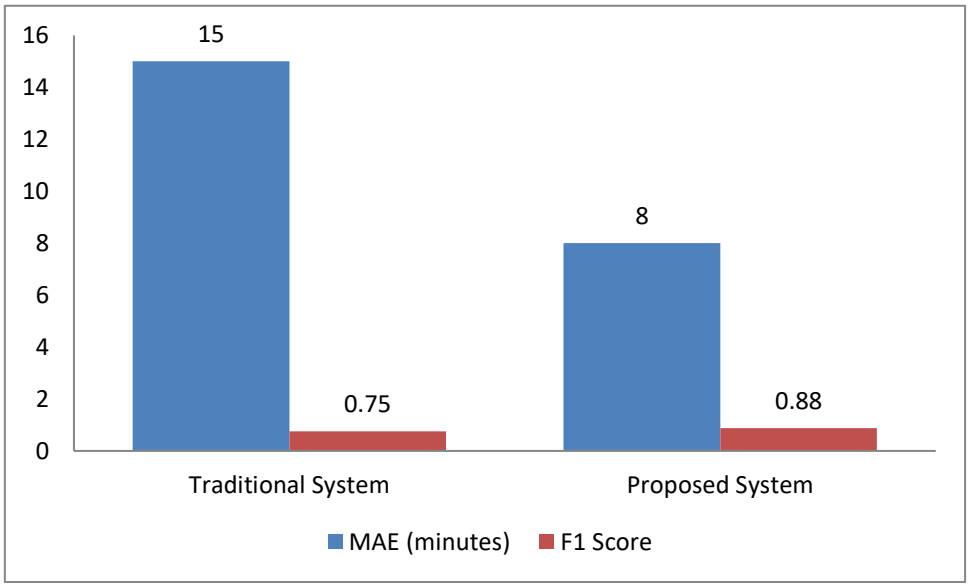

Figure 5: Graph representing Performance Metrics Comparison

4.2 Anomaly Detection Results

- **Traditional System:** Of all the sub-processes of business process monitoring, anomaly detection is vital in finding other patterns in work orders as well since high irregularities could mean errors, struggling systems or existing problems. The traditional anomaly detection system which was designed has an accuracy of 75%. This shows that the system was able to diagnose 75% of the work order anomalies, and there is still room to enhance the detection of other hidden or hard anxiety anomalies. Though such a degree of Precision may be useful for some scenarios, it raises false positives or can overlook an anomaly, which harms operational decisions. Secondly, sustaining such a level of accuracy with the increase in the number of work orders or the complexity of such work orders may cause operational risks, errors and inefficiencies.
- **Proposed System:** It was noticed that implementation of this proposed system yielded much better success in anomaly detection, attaining an accuracy rate of 92%. This improved accuracy is much higher than the previous system of doing the same work, by a difference of 17 percent. However, given such a high degree of accuracy, the proposed system has a better probability of identifying some of the flaws, which might otherwise not be easily noticeable, thereby enhancing general reliability as well as efficiency of work order management. The increase in accuracy decreases both false positive results and false negatives and allows for better identification of problems and their solving. This leads to improved anomaly detection, which is highly appreciated in complex environments with various and unpredictable disturbances as well as in situations where the system is able to adapt to the situation and quickly define the main difficulties affecting productivity and decreasing system available time.

Table 2: Anomaly Detection Results Comparison

System	Accuracy
Traditional System	75%
Proposed System	92%

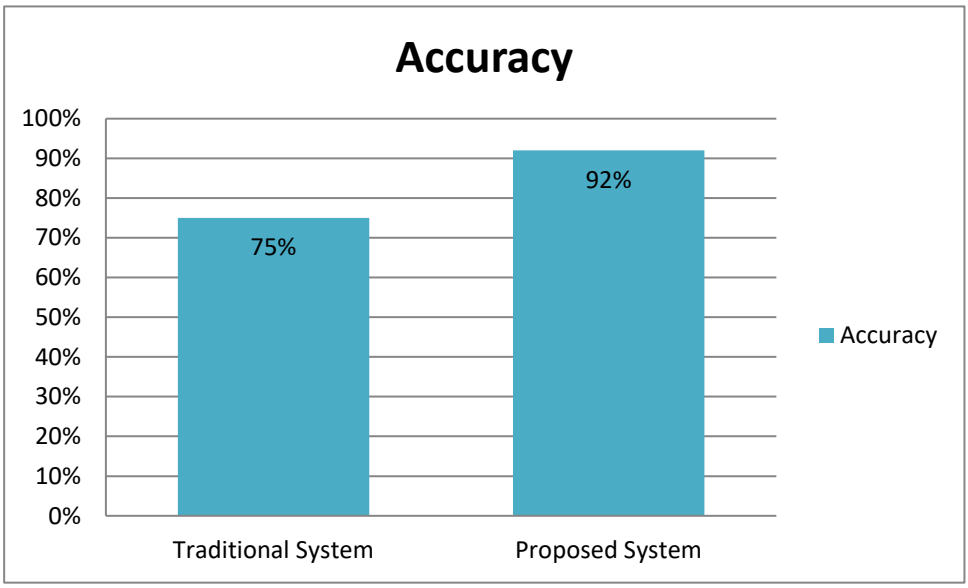

Figure 5: Graph representing Anomaly Detection Results Comparison

4.3. Discussion

4.3.1. Integration of Machine Learning (ML) and Databricks: The use of Machine Learning (ML) together with databricks was considered crucial to improving the work order management system. Integration of ML algorithms allowed performing two major tasks: considering big datasets and defining intricate patterns in the work order. Databricks helped in smooth data procession and developing and deploying large-scale machine learning models. These technologies enabled the system to cater to the growing 10,000 work orders dataset to get real-time information on task completion time and also use this as a dataset to predict anomalies. This scalability was of utter importance to the fact that the system remained accurate and efficient despite the increase in the amount of data. Further, one of the advantages of the system was the cross-sector flexibility, as the system itself could be easily adjusted to many different manufacturing environments. Hence, it supported the functionality of the enterprise and helped build better decision-making strategies.

4.3.2. Challenges

- **Data Privacy Concerns:** As the use of data-driven systems grew popular, data privacy became an issue that companies had to deal with. The work order management system deals with operational and employee information, information that has to be protected against access and misuse. Following compliance with data protection laws, including the GDPR, was a challenge since they introduced extra features into the system's development and deployment. To this end, the following key data security measures that would have to be put in place in the system were identified: use of data encryption, access controls and data anonymizing. These regulations were important not only because they are the law but also because failure to adhere to them would violate the trust that users and clients place in the system by providing access to their personal information.
- **Need for Domain Expertise:** It was revealed in the study that the machine learning models that have been incorporated into the system provided high accuracy, but the processes of designing and fine-tuning the models needed information from the domain specialists. These experts, endowed with adequate

manufacturing processes and operation constraints, ensured that models were well-tuned in operations and scenarios. Since a machine learning model depends on a number of assumptions, it is important that these assumptions are appropriate for the industry served; incorrect assumptions can vastly hamper a model's return. Hence, the engagement of data scientists together with domain experts was strategic in identifying how best to come up with a suitable system to find propounded formulas of job completion rates and review them correspondingly while recognizing irregular situations. The problem here was the trade-off between the level of skill needed to create the models and the level of domain knowledge necessary to make the system as useful as possible in real-world environments.

5. Conclusion

This study emphasizes the feature that combining the fields of ML with the Databricks platform has the potential to enhance the work order management of pharmaceutical manufacturing, which is a critical problem regarding efficiency, error, and legal compliance. In the current system design, the advanced equations of machine learning algorithms will be incorporated to improve the likelihood of correct estimation of tasks' completion time and identify anomalies. Using Databricks, a unified analytics platform, the system can accommodate big data and hence can be implemented in any manufacturing environment and provide real-time analysis. As a result of data processing and optimization, the business operating decisions are refined, relieving the potential for mistakes in tasks that are performed and optimising resource utilization.

In addition, ML deployed in work order management improves compliance with the specified regulatory guidelines, like those within the production of drugs. Where regulation is of the essence, it is important to keep records and ensure that the activities meet certain time frames. These needs are met in the proposed system since accuracy and traceability enable the pharmaceutical manufacturing processes to adhere to regulatory and industry requirements. However, the other benefit of the system is the probability of noncompliance by flagging outliers or irregularities; this reduces the risks of noncompliance to the bare minimum since the system can alert management of potential problems, such as quality or operation.

Further, in future research, efforts will be made to develop the methods and models used in the system as interpretable as possible. The current proposed model is already producing a high degree of accuracy and efficiency, but more research needs to be done to make the model more transparent and explainable for end-users and regulatory bodies. In general, interpretability will contribute to greater understanding and, therefore, acceptance of the system's operator choices. Furthermore, it is possible to implement blockchain technology for data confidentiality within the given system. In terms of recording and sharing data, blockchain technology can provide an additional layer of security and real-time updates to existing data. Plus, it's in line with the trend of patient safety in managed medication, such as those in the pharmaceutical industry. Because blockchain log entries are immutable, logs cannot be modified, which would help maintain the authenticity and confidentiality of data in manufacturing. Thus, by integrating these developments, the contemplated system will be more capable of supporting the industry, elaboration ng its further productivity increasmaintaininging the high quality he vital products, and adhering he constantly changing requirements.

References

1. Chowdary, B. V., & George, D. (2012). Improvement of manufacturing operations at a pharmaceutical company: a lean manufacturing approach. Journal of Manufacturing Technology Management, 23(1), 56-75.
2. Rybski, C., & Jochem, R. (2016). Benefits of a learning factory in the context of lean management for the pharmaceutical industry. Procedia CIRP, 54, 31-34.
3. Modgil, S., & Sharma, S. (2016). Total productive maintenance, total quality management and operational performance: An empirical study of Indian pharmaceutical industry. Journal of Quality in Maintenance Engineering, 22(4), 353-377.
4. Ding, B. (2018). Pharma Industry 4.0: Literature review and research opportunities in sustainable pharmaceutical supply chains. Process Safety and Environmental Protection, 119, 115-130.
5. Uthayakumar, R., & Priyan, S. (2013). Pharmaceutical supply chain and inventory management strategies: Optimization for a pharmaceutical company and a hospital. Operations Research for Health Care, 2(3), 52-64.
6. Deuel, A. C. (1994). The benefits of a manufacturing execution system for plantwide automation. ISA transactions, 33(2), 113-124.
7. Reska, D., Czajkowski, M., Jurczuk, K., Boldak, C., Kwedlo, W., Bauer, W., & Kretowski, M. (2021). Integration of solutions and services for multi-omics data analysis towards personalized medicine. biocybernetics and biomedical engineering, 41(4), 1646-1663.
8. Friedli, T., Goetzfried, M., & Basu, P. (2010). Analysis of the implementation of total productive maintenance, total quality management, and just-in-time in pharmaceutical manufacturing. Journal of Pharmaceutical Innovation, 5, 181-192.
9. Ganesh, S., Su, Q., Pepka, N., Rentz, B., Vann, L., Yazdanpanah, N., & Reklaitis, G. V. (2020). Design of condition-based maintenance framework for process operations management in pharmaceutical continuous manufacturing. International journal of pharmaceutics, 587, 119621.
10. Brown, S., & Vondráček, P. (2013). Implementing time-based manufacturing practices in pharmaceutical preparation manufacturers. Production Planning & Control, 24(1), 28-46.
11. Abideen, A. Z., & Mohamad, F. B. (2020). Supply chain lead time reduction in a pharmaceutical production warehouse–a case study. International Journal of Pharmaceutical and Healthcare Marketing, 14(1), 61-88.
12. Casola, G., Siegmund, C., Mattern, M., & Sugiyama, H. (2019). Data mining algorithm for preprocessing biopharmaceutical drug product manufacturing records. Computers & Chemical Engineering, 124, 253-269.
13. Chi, H. M., Moskowitz, H., Ersoy, O. K., Altinkemer, K., Gavin, P. F., Huff, B. E., & Olsen, B. A. (2009). Machine learning and genetic algorithms in pharmaceutical development and manufacturing processes. Decision Support Systems, 48(1), 69-80.
14. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity, 25, 1315-1360.
15. Nagaprasad, S., Padmaja, D. L., Qureshi, Y., Bangare, S. L., Mishra, M., & Mazumdar, B. D. (2021). Investigating the impact of machine learning in pharmaceutical industry. Journal of Pharmaceutical Research International, 33(46A), 6-14.
16. Selvaraj, C., Chandra, I., & Singh, S. K. (2021). Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. Molecular diversity, 1-21.
17. Djuriš, J., Kurćubić, I., & Ibrić, S. (2021). Review of machine learning algorithms application in pharmaceutical technology. Archives of Pharmacy, 71(Notebook 4), 302-317.

18. Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development Discussions, 2022, 1-10.
19. Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of Precision, recall, and f1 score for more thorough evaluation of classification models, in Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).
20. Zeberli, A., Badr, S., Siegmund, C., Mattern, M., & Sugiyama, H. (2021). Data-driven anomaly detection and diagnostics for changeover processes in biopharmaceutical drug product manufacturing. Chemical Engineering Research and Design, 167, 53-62.