

Research Article

Predictive Machine Learning Model for SAP Production Scheduling, Inventory Control, and Throughput Enhancement

Mai Phuong Hoang ¹

¹Center for Intelligent Machine Learning, Vietnam



Received: 25 June 2026
Revised: 20 July 2026
Accepted: 14 August 2026
Published: 01 September 2026
Doi:10.55640/ijdsml-06-02-06
Page No: 149-159

Copyright: © 2026 Authors retain the copyright of their manuscripts, and all Open Access articles are disseminated under the terms of the Creative Commons Attribution License 4.0 (CC-BY), which licenses unrestricted use, distribution, and reproduction in any medium, provided that the original work is appropriately cited.

Abstract

The increasing complexity of medication usage, drug misidentification, and dispensing-related errors has created a significant demand for intelligent systems capable of automated pharmaceutical recognition. Traditional approaches for identifying drugs rely heavily on manual inspection, prescription interpretation, and human expertise, which may introduce delays and inaccuracies. This research presents a machine learning-based framework for automated drug detection and visual recognition in self-images, designed to identify pharmaceutical products from user-captured images through advanced visual feature extraction, classification, and recognition mechanisms. The proposed framework integrates image preprocessing, object detection, feature representation, and machine learning-based classification to enhance medication identification efficiency. The study conceptually investigates how artificial intelligence can support safer medication practices by reducing recognition errors and improving accessibility to drug information. Existing research on medication errors, prescription recognition, and pharmaceutical image analysis highlights the need for automated solutions that combine computer vision with intelligent decision-making capabilities (Gates et al., 2019; Nayak et al., 2023). Furthermore, human-centered AI trust modeling principles emphasize the importance of reliable and interpretable intelligent systems when applied to critical domains such as healthcare (Ramamurthy et al., 2026). The framework addresses current limitations in manual drug identification by providing a scalable approach for real-world applications, including personal medication management, pharmacy assistance, and healthcare support systems. The research contributes a structured machine learning architecture for improving drug recognition accuracy, reducing medication-related risks, and strengthening intelligent healthcare technologies.

Keywords: Machine Learning, SAP Production Planning, Predictive Scheduling, Inventory Optimization, Throughput Enhancement, Production Analytics, Intelligent Manufacturing, Predictive Inventory Control, Lean Manufacturing

INTRODUCTION

1.1 Background

Medication Production planning in contemporary manufacturing requires simultaneous coordination of demand, material availability, inventory, production capacity, machine utilization, delivery commitments, and operational constraints. SAP-based enterprise planning environments provide structured mechanisms for managing these activities, but the quality of planning decisions remains strongly dependent on the accuracy of master data, planning parameters, historical information, and the ability of planners to anticipate future disruptions. A production schedule that appears feasible at the planning stage may

become inefficient when inventory shortages, unexpected capacity restrictions, delayed materials, demand changes, or production deviations occur.

Machine learning provides an alternative analytical mechanism because it can identify nonlinear relationships among large numbers of operational variables. Modern intelligent systems have demonstrated increasingly sophisticated capabilities for processing and reasoning over complex information. GPT-4 research, for example, illustrates the evolution of large-scale intelligent models capable of handling heterogeneous information and complex tasks (Achiam et al., 2023). Similarly, the development of Qwen and Falcon demonstrates the expanding diversity of machine-learning architectures and model configurations for intelligent information processing (Bai et al., 2023; Almazrouei et al., 2023). Although these models are not production-planning systems themselves, their development reinforces the broader theoretical premise that data-driven models can extract useful patterns from complex operational environments.

Within production planning, the central challenge is therefore not simply predicting demand or inventory independently. The more significant challenge is establishing a coordinated predictive mechanism in which scheduling, inventory control, and throughput are treated as interconnected decision variables. Lean-oriented SAP planning research emphasizes the value of aligning production planning with inventory and throughput objectives rather than optimizing isolated operational indicators (Kalal, 2026). This principle forms the conceptual foundation of the proposed model.

1.2 Problem Statement

Conventional production planning can suffer from three interconnected weaknesses. First, schedules may rely excessively on static parameters and historical averages, reducing responsiveness to emerging operational conditions. Second, inventory decisions may be separated from production scheduling, causing either excessive inventory accumulation or material shortages. Third, throughput optimization may occur after bottlenecks have already affected production rather than through predictive identification.

These weaknesses create a planning environment in which local optimization can generate system-level inefficiency. Increasing production for one product may improve utilization while simultaneously consuming materials required for higher-priority orders. Similarly, minimizing inventory may reduce carrying costs while increasing stockout risk and production interruptions. A machine-learning model must therefore account for relationships among these variables rather than treating each as an independent prediction problem.

1.3 Research Relevance

The research is relevant because SAP environments contain extensive operational data generated through production orders, material movements, inventory transactions, work-center activities, purchasing processes, and scheduling events. Transforming this historical information into predictive intelligence can enable planners to identify risks before they become operational problems.

The proposed approach also aligns with the broader evolution of intelligent models toward retrieval, contextual analysis, and adaptive processing. Retrieval-augmented language modeling has demonstrated the value of integrating external information with model reasoning (Borgeaud et al., 2022), while long-context transformer research illustrates the importance of handling extended information sequences (Bertsch et al., 2024). These developments provide theoretical support for designing production analytics systems capable of combining historical operational records with current SAP

planning conditions.

1.4 Objectives

The primary objective is to develop a predictive machine-learning framework capable of supporting SAP production scheduling, inventory control, and throughput enhancement simultaneously.

The specific objectives are to establish a structured SAP data representation; identify predictive variables associated with production delays and inventory risk; develop an integrated scheduling and throughput prediction mechanism; provide predictive alerts for material shortages and bottlenecks; and establish a feedback mechanism through which actual production outcomes can improve subsequent predictions.

1.5 Scope and Significance

The framework focuses on production-oriented SAP environments and addresses planning decisions rather than autonomous physical control. Its analytical scope includes production orders, material requirements, inventory states, production lead times, capacity utilization, scheduling deviations, and throughput indicators.

The significance of the framework lies in its integration of three traditionally interconnected but analytically separated objectives: schedule feasibility, inventory availability, and throughput efficiency. The proposed model therefore provides a decision-support layer rather than replacing SAP's transactional and planning infrastructure.

LITERATURE REVIEW

2.1 Intelligent Machine Learning as a Planning Foundation

Recent developments in intelligent machine learning demonstrate substantial progress in model capacity, contextual processing, and domain adaptation. Achiam et al. (2023) demonstrate the capabilities of large-scale language models for complex information processing, while Bai et al. (2023) and Almazrouei et al. (2023) illustrate the development of alternative large-model ecosystems. Bellagente et al. (2024) further demonstrate continued development of smaller language models, which is relevant to enterprise environments where computational efficiency and deployment constraints are important.

These developments are theoretically relevant to SAP analytics because production planning generates heterogeneous information at different temporal and organizational levels. A predictive planning system must process transactional histories, numerical measurements, categorical variables, temporal relationships, and operational constraints. The progression from large general-purpose models toward smaller and more deployable architectures suggests that intelligent analytics can increasingly be incorporated into enterprise environments without requiring every analytical task to depend on extremely large computational infrastructures.

2.2 Retrieval and Long-Context Processing

Production planning is inherently historical. Current scheduling decisions may depend on previous production performance, supplier lead times, inventory consumption, order priorities, and recurring bottlenecks. Borgeaud et al. (2022) demonstrate the importance of retrieving information from extensive external corpora, while Bertsch et al. (2024) investigate long-range transformer processing. These contributions support the conceptual argument that production intelligence should not rely exclusively on a narrow

recent data window.

For SAP planning, long-term historical patterns can reveal recurring production delays or material consumption anomalies. However, unlimited historical information is not automatically beneficial. Excessive historical data may introduce outdated patterns that are no longer representative of current production conditions. Consequently, a production-planning model requires temporal weighting, feature selection, and monitoring mechanisms rather than indiscriminate historical aggregation.

2.3 Model Efficiency and Enterprise Deployment

Enterprise machine-learning systems require a balance between predictive accuracy, computational cost, interpretability, and deployment feasibility. The Falcon series, Qwen, Phi-3, Stable LM, and Claude research collectively illustrate different approaches to model capability and deployment (Almazrouei et al., 2023; Bai et al., 2023; Abdin et al.; Bellagente et al., 2024; Anthropic AI, 2024). The relevance to SAP environments is primarily architectural: intelligent planning systems should be designed according to the operational problem rather than assuming that the largest model will necessarily produce the best business outcome.

For production scheduling, a compact gradient-based, tree-based, or hybrid predictive model may be preferable when the inputs are predominantly structured SAP data. Large language models may instead provide supplementary functions such as interpreting planner notes, summarizing exceptions, or assisting with natural-language explanations. This distinction prevents unnecessary complexity.

2.4 Knowledge Retrieval and Decision Context

Borgeaud et al. (2022) establish an important conceptual distinction between parametric knowledge and retrieved information. In production planning, an analogous distinction exists between information encoded through a trained predictive model and current operational information retrieved directly from SAP.

A predictive model trained on historical data cannot automatically know the current inventory position or latest production order. Therefore, real-time SAP information must remain authoritative for current-state variables, while machine learning should estimate future outcomes. This creates a hybrid architecture in which SAP provides operational truth and machine learning provides predictive intelligence.

2.5 Human Knowledge and Semantic Information

Bayley and Squire (2005) demonstrate the importance of semantic knowledge acquisition in cognitive systems. Although their study concerns human memory rather than industrial machine learning, it provides a useful theoretical perspective: effective decision-making requires meaningful representation of information rather than simple storage of isolated observations.

In SAP planning, a production order number without contextual interpretation provides limited analytical value. Features such as order priority, material criticality, historical delay frequency, capacity pressure, and supplier reliability transform raw transactions into decision-relevant information.

2.6 Research Gap

The provided literature largely addresses intelligent models, retrieval, long-context processing, and model development rather than integrated SAP production planning. This creates a clear application-level research gap. Existing intelligent-model research

establishes technical foundations for complex information processing, but it does not directly provide a unified mechanism for production scheduling, inventory control, and throughput optimization.

The lean-oriented SAP production-planning work of Kalal (2026) provides a more direct conceptual bridge by emphasizing the relationship between production planning, inventory, and throughput. The proposed framework extends this direction by introducing predictive machine learning as an analytical layer for anticipating operational conditions rather than relying solely on deterministic planning logic (Kalal, 2026).

METHODOLOGY

3.1 Research Design

This study adopts a design-oriented predictive analytics methodology. The objective is to construct an integrated machine-learning framework rather than evaluate a single algorithm in isolation. The proposed architecture contains six interconnected layers: SAP data acquisition, data preparation, feature engineering, predictive modeling, decision optimization, and feedback-based model updating.

The methodology treats production planning as a multi-objective prediction problem. Instead of generating one prediction, the system estimates several operational outcomes that influence one another. These include production-delay probability, inventory shortage risk, expected throughput, capacity pressure, and schedule feasibility.

3.2 SAP Data Acquisition Layer

The first layer collects structured information from SAP production and inventory processes. Representative variables include production order quantity, planned start date, planned completion date, actual completion date, material requirement, current stock, safety stock, open purchase quantities, supplier lead time, work-center capacity, machine utilization, historical production duration, and previous schedule deviations.

The objective is not merely to collect a large dataset but to establish a consistent analytical representation. Each production order can be represented as a temporal record containing material, capacity, inventory, and scheduling characteristics.

3.3 Data Preprocessing

SAP data may contain missing values, duplicate transactions, inconsistent timestamps, abnormal quantities, and changes in master-data structures. Preprocessing therefore involves missing-value treatment, duplicate identification, temporal alignment, categorical encoding, outlier assessment, and normalization where required.

Historical data should be divided chronologically into training, validation, and testing periods. Random partitioning can produce temporal leakage because future production conditions may unintentionally enter the training set. A chronological division better reflects real production deployment.

3.4 Feature Engineering

Feature engineering converts SAP transactions into predictive indicators. Important feature groups include:

Scheduling features: planned duration, historical schedule variance, order priority, sequence position, and production frequency.

Inventory features: current stock, safety-stock ratio, stock coverage, inventory consumption rate, open purchase quantity, and shortage frequency.

Capacity features: work-center utilization, available capacity, queue length, machine loading, and historical bottleneck frequency.

Throughput features: completed quantity per period, average cycle time, production downtime, order completion rate, and throughput variability.

A composite operational-risk representation can be expressed conceptually as:

$$R_t = w_s S_t + w_i I_t + w_c C_t + w_b B_t$$

where (S_t) represents scheduling risk, (I_t) inventory risk, (C_t) capacity pressure, and (B_t) bottleneck risk. The weights (w_s, w_i, w_c, w_b) can be estimated using validation performance or operational priorities.

3.5 Predictive Modeling Layer

The proposed architecture uses a hybrid predictive strategy. Structured machine-learning models can estimate production delay and inventory shortage probabilities, while regression or time-series components estimate throughput. Where sufficiently rich historical data exist, neural architectures can model complex temporal dependencies.

The prediction vector can be represented as:

$$\hat{Y}_{t+h} = [\hat{D}_{t+h}, \hat{I}_{t+h}, \hat{T}_{t+h}, \hat{C}_{t+h}]$$

where ($\hat{D}$) represents predicted production delay, ($\hat{I}$) inventory shortage risk, ($\hat{T}$) expected throughput, and ($\hat{C}$) expected capacity pressure.

The framework does not require a large language model for every prediction. Research on models such as GPT-4, Phi-3, Qwen, Falcon, and Stable LM demonstrates the diversity of intelligent architectures, but the appropriate model should depend on the structure and complexity of the production task (Achiam et al., 2023; Abdin et al.; Bai et al., 2023; Almazrouei et al., 2023; Bellagente et al., 2024).

3.6 Scheduling Decision Mechanism

Predictions are transformed into scheduling recommendations through a constraint-aware decision layer. A candidate production sequence can be evaluated according to expected completion time, material availability, capacity utilization, inventory risk, and throughput contribution.

A conceptual objective function is:

[

$$\min Z =$$

$$\alpha D + \beta I + \gamma C - \delta T$$

]

where (D) denotes schedule delay, (I) inventory risk, (C) capacity imbalance, and (T) throughput. The coefficients represent business priorities.

This structure prevents the model from maximizing throughput at the expense of inventory stability or delivery performance. Such integrated optimization is consistent with the lean-oriented principle that production planning should balance inventory and throughput rather than optimize one metric independently (Kalal, 2026).

3.7 Inventory Control Mechanism

The inventory component predicts the probability that available stock will become insufficient before the next replenishment opportunity. Instead of responding only after a stockout occurs, the system generates a predictive inventory-risk score.

For example, if historical consumption indicates that a critical material will fall below safety stock during a high-priority production period, the system can identify the conflict before the production schedule is finalized. The planner can then reschedule production, prioritize replenishment, or allocate available material to higher-value orders.

3.8 Throughput Enhancement Mechanism

Throughput enhancement focuses on identifying constraints that limit completed production quantity. The model evaluates work-center utilization, queue length, cycle-time variation, downtime patterns, and material availability.

A high-utilization work center is not automatically a bottleneck. If downstream capacity is available and cycle time remains stable, increased utilization may represent efficient operation. Conversely, moderate utilization accompanied by repeated delays may indicate another constraint, such as material shortages or setup inefficiency. Machine learning is therefore valuable because it can model interactions rather than relying on single-threshold rules.

3.9 Feedback and Continuous Learning

After production execution, actual outcomes are returned to the analytical layer. Predicted completion times are compared with actual completion times, predicted inventory risks with actual shortages, and expected throughput with realized throughput.

Prediction error can be represented as:

[

$$E_t = Y_t - \hat{Y}_t$$

]

Persistent errors indicate potential concept drift, changed production conditions, master-data problems, or inadequate feature representation. Continuous monitoring is therefore essential for maintaining model reliability.

RESULTS

The proposed framework produces several analytical findings concerning the integration of predictive machine learning with SAP production planning. First, production scheduling should be treated as a predictive decision problem rather than solely as a deterministic sequencing activity. Historical schedule deviations, production duration, capacity pressure, material availability, and order characteristics provide signals that can be transformed into an estimate of future schedule risk. This enables planners to identify vulnerable orders before planned completion dates are missed. The finding is consistent with the broader movement toward increasingly capable intelligent models that extract relationships from complex information rather than depending exclusively on predefined rules (Achiam et al., 2023).

Second, inventory control and production scheduling are analytically inseparable. A schedule can be technically feasible in terms of machine capacity while remaining operationally infeasible because required materials are unavailable. The proposed framework addresses this problem by incorporating inventory-risk predictions into schedule evaluation. Instead of treating inventory as a static quantity, the model considers consumption rate, replenishment timing, safety-stock conditions, and production requirements. This creates a forward-looking inventory mechanism capable of identifying potential shortages before they disrupt production.

Third, throughput enhancement benefits from combining capacity information with production and inventory signals. A machine or work center with high utilization does not necessarily represent the primary constraint. The proposed model therefore evaluates utilization jointly with queue length, cycle time, material availability, and historical completion behavior. This supports earlier identification of bottlenecks and reduces the risk of optimizing one resource while transferring constraints elsewhere.

Fourth, the framework indicates that predictive intelligence should complement rather than replace SAP planning logic. SAP remains the authoritative environment for transactional and operational information, while machine learning estimates future conditions. This distinction is important because historical models cannot reliably represent current operational states unless they are continuously updated with current SAP data. Research on retrieval-based systems provides a broader theoretical justification for combining learned representations with dynamically retrieved information (Borgeaud et al., 2022).

Fifth, model size alone is not a sufficient indicator of enterprise planning effectiveness. Research covering GPT-4, Qwen, Falcon, Phi-3, Claude, and Stable LM demonstrates substantial variation in model capability and deployment characteristics (Achiam et al., 2023; Bai et al., 2023; Almazrouei et al., 2023; Abdin et al.; Anthropic AI, 2024; Bellagente et al., 2024). For structured SAP planning data, a smaller specialized predictive model may offer superior operational practicality because it can be easier to validate, explain, maintain, and integrate.

Finally, the integrated architecture provides a stronger theoretical basis for lean production planning than isolated prediction models. By simultaneously considering schedule stability, inventory risk, capacity pressure, and throughput, the framework operationalizes the principle of balancing competing production objectives (Kalal, 2026). The principal finding is therefore architectural: value emerges not from prediction alone but from connecting predictions to constrained planning decisions and subsequent execution feedback.

DISCUSSION

The findings indicate that machine learning can strengthen SAP production planning primarily by improving anticipation rather than by replacing established planning

mechanisms. This distinction has theoretical importance. Traditional planning systems generally determine feasible actions from known parameters and constraints, whereas predictive analytics estimates how current conditions are likely to evolve. Combining both approaches produces a hybrid decision architecture in which SAP establishes operational state and constraints while machine learning estimates future risk.

The integration of inventory and scheduling represents one of the most significant contributions of the proposed model. In isolated planning environments, inventory optimization can encourage lower stock levels while scheduling optimization can prioritize high utilization. These objectives may conflict. A production schedule that maximizes utilization can consume scarce material and increase future shortage risk. Conversely, aggressively minimizing inventory can increase production interruptions. The proposed multi-objective representation addresses this contradiction by evaluating inventory, scheduling, capacity, and throughput jointly. This corresponds closely with lean-oriented SAP production optimization principles described by Kalal (2026).

The framework also has implications for intelligent enterprise architecture. Contemporary model research increasingly emphasizes retrieval, long-context processing, and model specialization (Borgeaud et al., 2022; Bertsch et al., 2024). These developments suggest that an enterprise planning system should not be designed as a single monolithic model. A more robust architecture separates structured prediction from contextual interpretation. Numerical machine-learning models can predict delay and inventory risk, while language models may assist with exception explanations, planner communication, and interpretation of unstructured notes. This separation can reduce unnecessary computational complexity while preserving the benefits of intelligent interfaces.

A further implication concerns explainability. Production planners are unlikely to accept scheduling recommendations merely because a model assigns them a high probability. They need to understand why a particular order has been identified as risky. The proposed feature-based architecture supports explanations through variables such as insufficient material coverage, high work-center utilization, abnormal cycle time, or historical schedule variance. Explainability therefore becomes not only a technical property but also an organizational requirement.

Several limitations remain. The proposed model is conceptual and requires validation using actual SAP production datasets before quantitative claims about accuracy, cost reduction, or throughput improvement can be established. Model performance may also deteriorate when manufacturing processes, product portfolios, supplier conditions, or planning policies change. Historical relationships are not necessarily stable over time. Data quality is another major limitation because incorrect material masters, missing production confirmations, or inconsistent timestamps can produce misleading predictions.

The literature supplied for this study also creates a domain limitation. Most provided references concern language models, retrieval, knowledge processing, or intelligent-model development rather than SAP production planning specifically. Consequently, their role is primarily theoretical and architectural rather than empirical evidence for manufacturing outcomes. The direct SAP-oriented contribution of Kalal (2026) is therefore particularly important because it provides the conceptual connection between lean planning, inventory, and throughput.

The practical implication is that organizations should introduce predictive machine learning incrementally. A production environment can initially deploy delay-risk prediction and inventory alerts before progressing toward automated scheduling recommendations. Human planners should remain involved in high-impact decisions,

particularly when predictions conflict with contractual priorities, engineering constraints, or exceptional operational circumstances.

CONCLUSION

This research proposed a predictive machine-learning model for integrating SAP production scheduling, inventory control, and throughput enhancement. The framework addresses a fundamental weakness of isolated planning approaches by treating production schedules, inventory availability, capacity pressure, and throughput as interdependent variables.

The proposed architecture establishes six principal layers: SAP data acquisition, preprocessing, feature engineering, predictive modeling, decision optimization, and continuous feedback. Its central contribution is the transformation of historical and current SAP information into predictive indicators that can support earlier identification of schedule delays, material shortages, capacity constraints, and throughput limitations.

The analysis also demonstrates that machine learning should complement SAP rather than replace it. SAP provides authoritative transactional and operational information, while predictive models estimate future conditions. This hybrid relationship enables planners to retain business constraints and operational control while gaining stronger predictive capabilities. The approach is consistent with lean-oriented production planning principles that emphasize balanced optimization of inventory and throughput (Kalal, 2026).

Future research should empirically validate the framework using longitudinal SAP production datasets, compare alternative machine-learning algorithms, evaluate prediction drift, quantify inventory and throughput improvements, and investigate explainable AI mechanisms for planner acceptance. Further work may also explore specialized language models for interpreting unstructured production information while retaining structured predictive models for numerical planning tasks. Such research could advance SAP production planning from reactive transaction management toward continuously adaptive, predictive, and evidence-based decision support.

REFERENCES

1. Chowdhury, 1. M. Abdin et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv:
2. J. Achiam et al., GPT-4 technical report, arXiv:2303.08774, 2023.
3. J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, GQA: Training generalized multi-query transformer models from multi-head checkpoints, arXiv:2305.13245, 2023.
4. Z. Allen-Zhu and Y. Li, Physics of language models: Part 3.3, knowledge capacity scaling laws, arXiv:
5. E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, The falcon series of open language models, arXiv:2311.16867, 2023.
6. Anthropic AI, The Claude 3 model family: Opus, Sonnet, Haiku, Claude-3 Model Card <https://www.anthropic.com/claude>, 2024.
7. Argilla, Distilabel Math Preference DPO. <https://huggingface.co/datasets/argilla/distilabelmath-preference-dpo>, 2023.
8. Azure AI Services. GPT-4 and GPT-4 turbo models. <https://learn.microsoft.com/en-us/azure/aiservices/openai/concepts/models#gpt-4-and-gpt-4-turbo-models>, 2024. [Accessed 22-05-2024]

9. J. Bai et al., Qwen technical report, arXiv:2309.16609, 2023.
10. P. J. Bayley and L. R. Squire, Failure to acquire new semantic knowledge in patients with large medial temporal lobe lesions, *Hippocampus*, 15(2):273–280, 2005.
11. E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf, Open LLM leaderboard, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard,
12. M. Bellagente et al., Stable lm 2 1.6 b technical report, arXiv:2402.17834, 2024.
13. A. Bertsch, U. Alon, G. Neubig, and M. Gormley, Unlimiformer: Long-range transformers with unlimited length input, *Adv. Neural Inf. Process. Syst.*, Vol. 36, 2024.
14. S. Borgeaud et al., Improving language models by retrieving from trillions of tokens, in: *International conference on machine learning*, PMLR, 2206–2240, 2022.
15. M. Kalal, "Lean-Driven SAP Production Planning Optimization Using Machine Learning for Inventory and Throughput Efficiency," 2026 14th International Symposium on Digital Forensics and Security (ISDFS), Boston, MA, USA, 2026, pp. 1-6, doi: 10.1109/ISDFS69419.2026.11459029.