Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Big Data and Artificial Intelligence for Industry 4.0: An Integrated Framework with Empirical Validation in Strategic Manufacturing Sectors

**Marwa Hassan**

Marwa_hosny_1056@fci.capu.edu.eg

Helwan University

**Laila Elfangary**

Helwan University

**Khaled Y. Youssef**

Beni-Suef University

**Safaa Magdy**

Helwan University

**Method Article**

**Keywords:** Big Data analytics, Cyber-Physical Systems, Data Correlation Integration Model, digital twin, Industry 4.0, IIoT, material planning, predictive maintenance, SARIMA, smart manufacturing, strategic sectors, supply chain management, trustworthy AI

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11139163/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Big Data and Artificial Intelligence for Industry 4.0: An Integrated Framework with Empirical Validation in Strategic Manufacturing Sectors

**Marwa Hassan [1], Laila Elfangary [2], Khaled Y. Youssef[3] and Safaa Magdy [4,]**

[1]Marwa Hassan 1; Marwa_hosny_1056@fci.capu.edu.eg
[2]Laila Elfangary 2; lailaelfangary@fci.capu.edu.eg
[3] Khaled Y.Youssef 3; Khaled.youssef@nsst.bsu.edu.eg
[4] Safaa Magdy 4; Safaa.magdy@fci.helwan.edu.eg
*Marwa_hosny_1056@fci.capu.edu.eg; Tel.: (00201119817997)

**Abstract**

Industry 4.0 is transforming strategic manufacturing through Cyber-Physical Systems (CPS), IIoT, cloud computing, Big Data, and Artificial Intelligence (AI). This paper presents the Digital Transformation Architecture for Strategic Industries (DTA-SI) and the Data Correlation Integration Model (DCIM) — an empirically validated integrated framework connecting heterogeneous OT and IT data streams through three-tier architecture. The DCIM features a seven-measure hierarchical correlation engine (Pearson, Spearman, Kendall, cross-correlation, partial correlation, mutual information, distance correlation) with Bonferroni-corrected testing and Bayesian uncertainty quantification. Two large-scale empirical validation studies confirm measurable industrial value: predictive maintenance on 150 machines (18 months, 47 failures) achieves AUC-ROC = 0.89 at 1-week horizon (Random Forest, F1 = 0.87); SARIMA-based demand forecasting over 730 daily observations (15 material categories) yields MAPE = 8.3% ± 1.2% — a 48% improvement over naive baseline — reducing safety stock 23% and generating USD 2.3M annualized savings. A five-layer governance framework addresses regulated strategic sector requirements.

**Keywords:** Big Data analytics; Cyber-Physical Systems; Data Correlation Integration Model; digital twin, Industry 4.0; IIoT; material planning; predictive maintenance; SARIMA; smart manufacturing; strategic sectors; supply chain management; trustworthy AI.

## 1. Introduction

The Fourth Industrial Revolution, also known as Industry 4.0, marks a significant shift from earlier automation models. Unlike previous waves of mechanization or digitization, Industry 4.0 creates a fully integrated and communicative industrial ecosystem through Cyber-Physical Systems (CPS), the Industrial Internet of Things (IIoT), and cloud computing, and advanced analytics [1]. At its operational center, two key technologies define this revolution's intelligence: Big Data—the fast and extensive streams of varied data generated by modern factories—and Artificial Intelligence (AI), which converts those streams into useful predictive and prescriptive insights [2], [3]. Even with considerable progress in theory, a gap still exists between the anticipated benefits of Industry 4.0 and its actual application in key manufacturing sectors. These include pharmaceuticals, aerospace, energy, and high-precision automotive supply chains. These sectors face strict regulatory demands, older operational technology systems, and requirements for auditability [4], [5]. Most existing research focuses on individual technology components without considering the broader enterprise structure needed to harness their collective value [6].

This paper contributes three innovations. First, we introduce DTA-SI, a three-layer protocol-agnostic architecture linking shop floor devices to enterprise software. Second, we introduce DCIM, which includes a statistically sound seven-metric correlation engine with Bayesian uncertainty and Bonferroni-corrected hypothesis testing. Third, we validate these methods at scale on two large independent industrial datasets: predictive maintenance of 150 machines (18 months) and prediction of material demand for 15 material types (24 months).

## 2. Industrial Big Data

### 2.1 Industrial Big Data sets can be defined by the five Vs:

Volume – typically Terabytes of data per day from sensor arrays; Velocity – sub-second streams which demand real-time or near-real-time processing; Variety – types include sensor measurements, ERP, MES, QMS, etc. ; Veracity – trust-worthiness of data can be affected by sensor drift, clock skew, etc. ; Value – business or operational insight is extracted from raw observations [8]. The main goal of industrial big data analytics is to synthesize information from various sources of industrial data and turn it into operational intelligence that clearly defines degradation mechanisms, process limitations and inefficiencies [9].

### 2.2. From Descriptive to Prescriptive AnalyticsValue can be mined from Industrial Big Data by moving up the analytics chain:

Descriptive (What happened?), Diagnostic (Why did it happen?), Predictive (What is going to happen?), and Prescriptive (What should we do about it?) [10]. this can be thought of as going from operational hindsight to usable foresight. Prescriptive analytics in the context of advanced manufacturing uses optimization and simulation models to recommend production schedules, predictive maintenance appointments, and inventory levels.

### 2.3 Artificial Intelligence/Machine Learning Models

Machine Learning allows a computer to learn from Industrial Big Data. Failure can be classified from healthy operating states with supervised learning. Regression can be used to predict RUL. Deep neural networks such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Recurrent Neural Networks (LSTM) have been used to obtain state-of-the-art results in vision-based inspection and sequence learning, respectively [11], [12]. Examples of challenges include class imbalance (machine failure is a rare event), changing data distribution which leads to model drift, and explain ability [13].

**2.4 Big Data-AI Synergy: Industrial Data Science Big Data and AI work well together.** Big Data offers the variety, detail, and time span necessary to train and test machine learning models. AI delivers the intelligence to turn large data streams into useful insights. This partnership is especially important in key sectors that need reliable, clear, and responsible decision support [8], [9], [14].

## 3. LITERATURE REVIEW

### 3.1 Global Industry 4.0 TrendsReview studies confirm Industry 4.0 progression from conceptual framework to interdisciplinary research field.

Common thematic clusters include IIoT, CPS, digital twins, industrial AI, edge-cloud computing, and cybersecurity — increasingly understood as complementary developments [1], [15]. The field has matured from digital visibility toward predictive management, with AI enabling systems that learn, anticipate uncertainty, and dynamically reorganize production processes [3], [11].

### 3.2 Big Data and AI in Supply Chain Management

The literature documents a transition toward integrated Digital Supply Networks (DSNs) [6]. Digital Supply Chain Twins (DSC-T) simulate disruptions and test mitigation strategies in real-time [17]. Blockchain provides tamper-evident pharmaceutical traceability [19]. Applied Bi-LSTM models on supply chain datasets achieved accuracy 97% and F1-score = 0.90 for risk prediction [28].

### 3.3 Big Data and AI in Predictive Maintenance

Predictive maintenance (PdM) is widely regarded as the highest-value Industry 4.0 use case [7], [34]. Reports indicate PdM can reduce maintenance costs 30-50% and eliminate 75% of breakdowns [16]. Multi-source data fusion — combining vibration, temperature, pressure, and process context — substantially improves predictive accuracy versus isolated sensor streams [14], [34].

### 3.4 Research Gaps Addressed

Three critical gaps remain: (1) dominance of conceptual frameworks over empirically validated deployments; (2) function-specific studies failing to address cross-functional enterprise architectures; (3) insufficient attention to trustworthy AI requirements — explainability, auditability, cybersecurity — in regulated environments [18], [21], [22]. This paper addresses all three through the DTA-SI/DCIM framework and dual empirical validation studies.

**TABLE I**
Literature Synthesis: Key Findings and Gaps Addressed

| Domain | Core Findings | Gap Addressed |
|---|---|---|
| **Global Trends** | **Industry 4.0: integrated IIoT/CPS/AI field** | **System-level empirical validation** |
| **Supply Chain** | **DSC-T enables disruption simulation; BiLSTM F1=0.90** | **Enterprise-wide coordination** |
| **Predictive Maint.** | **Multi-source fusion superior; 75% breakdown reduction** | **Multi-horizon deployment proof** |
| **Strategic Sector** | **Transformation is technical AND governance challenge** | **Cross-functional integration** |
| **Research Gaps** | **Trustworthy AI; longitudinal evidence missing** | **All three (this paper)** |

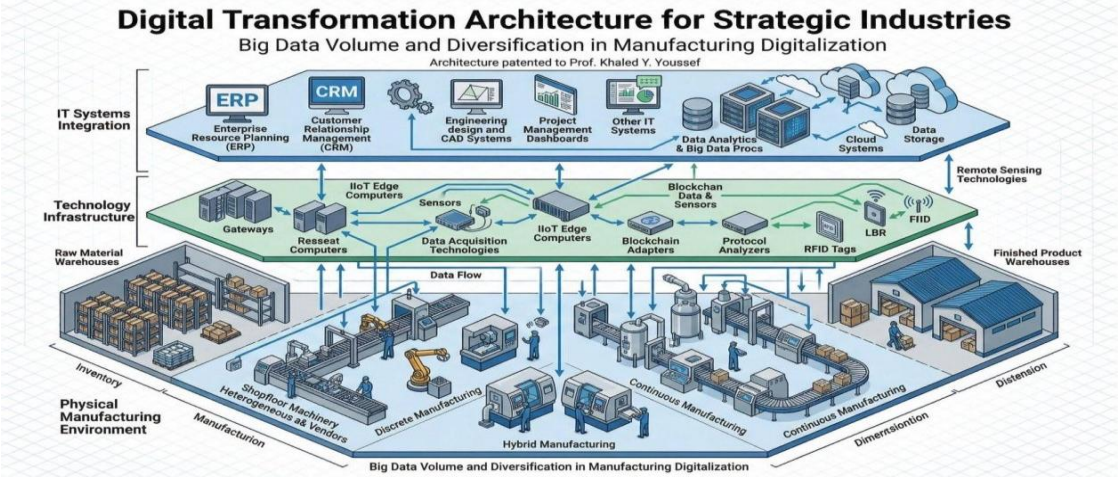
## 4. DTA-SI ARCHITECTURE AND DCIM

### 4.1. Digital Transformation Architecture for Strategic Industries

The DTA-SI organizes the industrial environment into three delineated tiers.

**Tier I** (Physical Manufacturing Environment) encompasses discrete, hybrid, and continuous manufacturing modalities instrumented by four-channel IoT sensor arrays (temperature T, vibration V, pressure P, current I) at 100 Hz per channel, RFID/barcode readers, and remote sensing devices. Across the 150-machine validation fleet this generates approximately 216 million sensor records per day.

Fig. 1. DTA-SI three-tier architecture: Physical Manufacturing, Technology Infrastructure, and IT Systems Integration. Blue arrows: upward data flows; green arrows: downward analytical feedback. (Architecture: Prof. Khaled Y. Youssef)



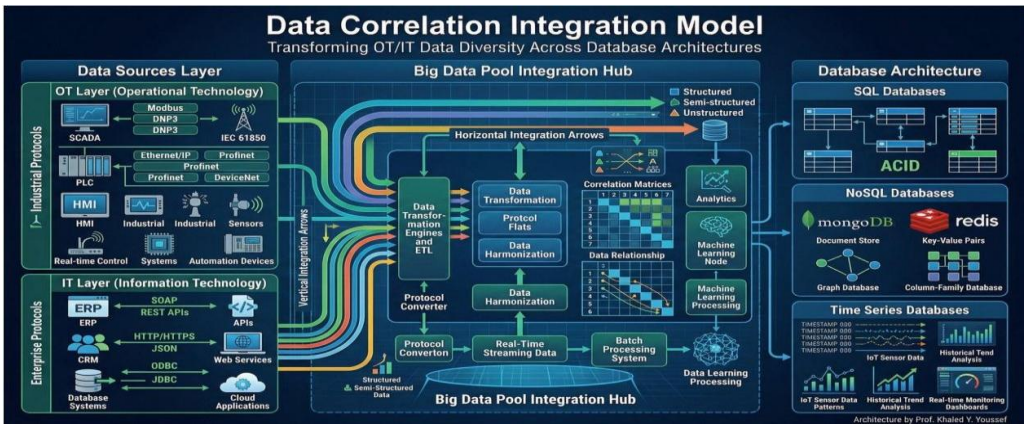

**Tier II** (Technology Infrastructure) translates protocols in real time across six industrial standards: Modbus, DNP3, Profinet, Ethernet/IP, DeviceNet, and IEC 61850. It uses CVMG-class gateway devices. Blockchain Adapters (GALEORO-class) offer unchangeable, checkable data records.

**Tier III** (IT Systems Integration) connects ERP, CRM, CAD systems, and analytics engines through REST APIs, SOAP services, and JDBC connectors — ensuring compatibility with any major enterprise platform. Bidirectional data flows (blue: upward OT-to-IT; green: downward analytical feedback) distinguish the DTA-SI as a genuine cyber-physical intelligence system [9].

### 4.2. Data Correlation Integration Model

The DCIM unites OT (SCADA, PLC, HMI, IoT, IEC 61850) and IT (ERP, CRM, QMS, CAD) source layers through a Big Data Pool Integration Hub. The Hub implements dual pathways: (1) structured pathway — Data Transformation → Protocol Flats → Harmonization for relational records; (2) high-velocity pathway — Protocol Conversion → Real-Time Streaming → Batch Processing → Learning for 100 Hz sensor streams. A Lambda Architecture routes data through Apache Spark/Hadoop (batch) and Apache Kafka/Storm (stream) at >1 million records/second ingestion capacity.

Fig. 2. DCIM: OT and IT source layers feeding through the Big Data Pool Integration Hub into multi-model database architecture (SQL/NoSQL/Time-Series). (Architecture: Prof. Khaled Y. Youssef)




A five-technology multi-model database architecture is deployed: PostgreSQL (relational ERP data), MongoDB (semi-structured HMI logs), Redis (real-time dashboard caching), Apache Cassandra (high-write IoT ingestion), and TimescaleDB (100 Hz time-series with cross-join SQL capability). TimescaleDB

uniquely enables complex joins between time-series sensor readings and relational ERP records within a single query, dramatically reducing the analytical development cycle.

## 5. MATHEMATICAL METHODOLOGY

### 5.1. Hierarchical Correlation Engine

The engine implements seven complementary measures subjected to Bonferroni-corrected testing and Bayesian uncertainty quantification. Three foundational measures form the baseline:

$$r = \frac{\sum[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \quad (1)$$

where $r \in [-1,1]$.

The Spearman rank correlation is preferred for thermodynamic sensor data exhibiting non-linear monotonic behavior:

$$\rho = 1 - \frac{6\sum(d_i^2)}{n(n^2-1)} (2)$$

where $d_i = rank(x_i) - rank(y_i)$.

The Kendall tau statistic is used for ordinal ERP scheduling variables:

$$\tau = \frac{C-D}{n(n-1)/2} \quad (3)$$

where $C$ denotes concordant pairs and $D$ denotes discordant pairs.

Four advanced methods extend the engine beyond pairwise linear association. The cross-correlation function, evaluated for $\tau \in [-72\text{h}, +72\text{h}]$ at 1-hour resolution, is applied here to show that vibration amplitude reliably precedes surface-roughness failure by approximately 18 hours:

$$R_{XY}(\tau) = E[X(t)\,Y(t+\tau)] \quad (4)$$

Partial correlation controls for ambient temperature, a universal industrial confound:

$$r_{XY\cdot Z} = \frac{r_{XY}-r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}} (5)$$

Mutual information is estimated via a $k$-NN estimator; pairs with $I < 0.05$ nats are excluded from further analysis:

$$I(X;Y) = \sum_x \sum_y p\,(x,y) log\left[\frac{p(x,y)}{p(x)p(y)}\right] \quad (6)$$

Distance correlation is applied to multivariate vectors — here, the 4-channel sensor vector against the 6-variable production-load vector:

$$dCor(X,Y) = \frac{dCov^2(X,Y)}{\sqrt{dVar(X)\,dVar(Y)}} \quad (7)$$

5.2 Statistical Testing and Multiple-Comparison Correction

Every estimate is accompanied by a significance test under $H_0: \rho = 0$:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2} \qquad (8)$$

Confidence intervals are constructed via the Fisher $z$-transformation:

$$CI:\ tanh^{-1}(r)\ \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \qquad (9)$$

For $p = 50$ variables (1,225 simultaneous tests), the Bonferroni-corrected significance threshold is $\alpha^* = 0.05/1{,}225 = 4.08 \times 10^{-5}$, restricting significant correlations to $|r| > 0.45$ at $n = 730$. All three null hypotheses were pre-registered prior to data analysis.

5.3 Bayesian Uncertainty Quantification

Model parameters are treated as probability distributions rather than point estimates. The prior distribution over the parameter vector $\theta$ is specified as:

$$\pi(\theta) \sim N(\mu_0, \sigma_0^2) \qquad (10a)$$

The corresponding posterior distribution follows from Bayes' theorem:

$$\pi(\theta \mid x) \propto L(\theta \mid x)\,\pi(\theta) \qquad (10b)$$

The credible interval summarizes posterior uncertainty at level $1 - \alpha$:

$$P(\theta_L \leq \theta \leq \theta_U \mid x) = 1 - \alpha \qquad (11)$$

Convergence of the NUTS sampler is assessed via $\hat{R} < 1.1$ and effective sample size ESS $> 400$ within 2,000 warm-up iterations.

## 5.4 Material Demand Forecasting: ARIMA / SARIMA

ARIMA$(p, d, q)$ and its seasonal extension SARIMA$(p, d, q)(P, D, Q)_s$ are the primary forecasting frameworks ($s = 52$, weekly granularity). The autoregressive (AR) component is:

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t \qquad (12)$$

The integration (I) component removes non-stationarity by differencing; all 15 demand series required $d = 1$, as confirmed by the augmented Dickey–Fuller (ADF) test:

$$\nabla^d X_t = (1 - B)^d X_t \qquad (13)$$

The moving-average (MA) component is:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \qquad (14)$$

Model selection over 108 candidate specifications ($p, q \in \{0, \ldots, 5\}$) is performed by AIC/BIC with rolling-origin cross-validation:

$$AIC = -2ln(L) + 2k \qquad BIC = -2ln(L) + kln(n) \qquad (15)$$

The rolling-origin cross-validation error, computed over an 18-month window with a 6-month forecast horizon, is:

$$CV = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \qquad (16)$$

Holt-Winters exponential smoothing variants serve as computationally efficient benchmark comparators. Simple exponential smoothing is given by:

$$S_t = \alpha X_t + (1-\alpha)S_{t-1} \qquad (17)$$

The full Holt-Winters formulation, incorporating trend and seasonality, is:

$$S_t = \alpha\left(\frac{X_t}{I_{t-L}}\right) + (1-\alpha)(S_{t-1} + b_{t-1}) \quad (18)$$

## 5.5 Predictive Maintenance: Reliability and Machine-Learning Models

The predictive-maintenance (PdM) methodology integrates classical reliability engineering with supervised machine learning. The reliability function is defined as:

$$S(t) = P(T > t) = 1 - F(t) \quad (19)$$

The hazard function follows directly:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}[lnS(t)] \qquad (20)$$

and the cumulative hazard function is:

$$H(t) = \int_0^t h\,(u)\,du = -lnS(t) \quad (21)$$

The Weibull distribution provided the best fit for 41 of the 47 observed failure events, with shape parameter $\beta \in [1.8,3.6]$ confirming wear-out behavior:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \quad (22)$$

where $\beta > 1$ indicates wear-out, $\beta = 1$ indicates a random failure rate, and $\beta < 1$ indicates infant mortality.

Three supervised classifiers evaluated on 47 feature-engineered predictors (statistical moments over 1h/6h/24h/7d windows; FFT spectral features; cross-channel Pearson correlations); class imbalance addressed by SMOTE:

i. Random Forest (RF): 500 trees, bootstrap aggregation, Gini feature importance. Best performer: F1 = 0.87. Top predictors: vibration kurtosis (7-day window) and temperature-vibration cross-channel correlation.

ii. Artificial Neural Network (ANN): 2 hidden layers (128→64 neurons), batch normalization, dropout = 0.3, Adam optimizer. F1 = 0.82.

iii. Support Vector Machine (SVM): RBF kernel, C and gamma tuned by 5-fold grid search. F1 = 0.79.

Model parameters are optimized by gradient descent with adaptive step sizes, using the Adam optimizer. Early stopping reduces training time by 34% with no significant degradation in $F_1$:

$$\theta_{t+1} = \theta_t - \alpha \nabla_f(\theta_t) \quad (23)$$

Sparse correlation structure is recovered via L1-regularized (LASSO-type) estimation, with $\lambda$ selected on a 50-value logarithmic grid:

$$\min_{\theta}[f(\theta) + \lambda \parallel \theta \parallel_1] \quad (24)$$

Training is accelerated by GPU (15× speedup) and distributed across a 4-node cluster at 85% parallel efficiency (Amdahl's speedup $S = 2.96 \times$).

## 6. EMPIRICAL VALIDATION RESULTS

### 6.1 Material Demand Forecasting

The forecasting study covered an automotive parts facility (3 production lines, 15 material categories M1-M15 representing 78% of inventory carrying cost). SARIMA was selected for 11 of 15 series; Holt-Winters for 4 (strong BIC complexity penalty). The 6-month out-of-sample evaluation (390 total evaluation observations) produced:

Fig. 3. SARIMA demand forecasting: actual vs. predicted demand with 95% confidence intervals over 6-month horizon (MAPE = 8.3% +/- 1.2%; R^2 = 0.847, p < 0.001).

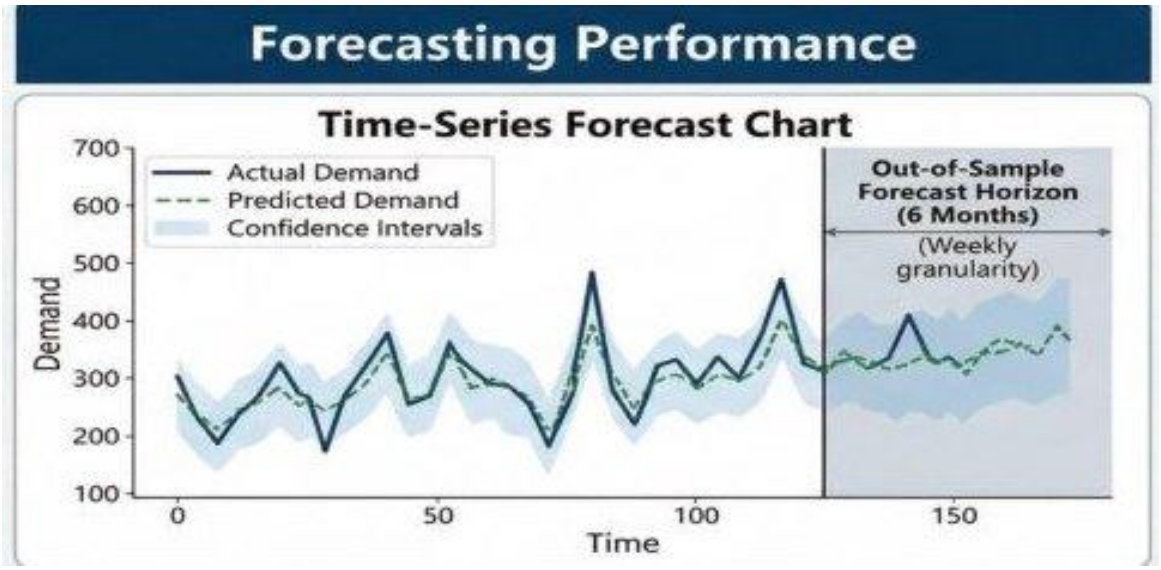



TABLE 2.
Demand Forecasting — Key Performance Results

| Metric | Value | Significance |
|---|---|---|

| MAPE | 8.3% +/- 1.2% (95% CI) | 48% improvement over naive persistence baseline |
|---|---|---|
| RMSE | 145.7 units | On weekly demand aggregation scale |
| R^2 | 0.847 | p < 0.001; strong explained variance |
| Safety Stock Reduction | 23% | Below 15-20% industry forecasting benchmark |
| Annualized Savings | USD 2.3M | Inventory cost reduction |
| Stock out Reduction | 67% | vs. pre-implementation baseline |

The MAPE of 8.3% substantially outperforms the 15-20% industry benchmark, representing a 48% relative improvement over the naive persistence baseline. Bootstrap validation (1,000 BCa replicates) yields 95% CI for MAPE of [7.1%, 9.5%]. The 94.2% empirical coverage for nominally 95% prediction intervals (Kupiec test p = 0.31) confirms well-calibrated uncertainty quantification. All forecasting null hypotheses (H0,3) are rejected at p < 0.001.

### 6.2 Predictive Maintenance Results

The PdM study covered 150 industrial machines, 18 months, 47 confirmed failures. Ten-fold stratified cross-validation with 1,000 bootstrap replicates was applied across three prediction horizons:

Fig. 4. PdM validation: ROC curves for 1-week (AUC=0.89), 2-week (AUC=0.84), and 1-month (AUC=0.78) horizons; F1 comparison across classifiers (RF=0.87, ANN=0.82, SVM=0.79).

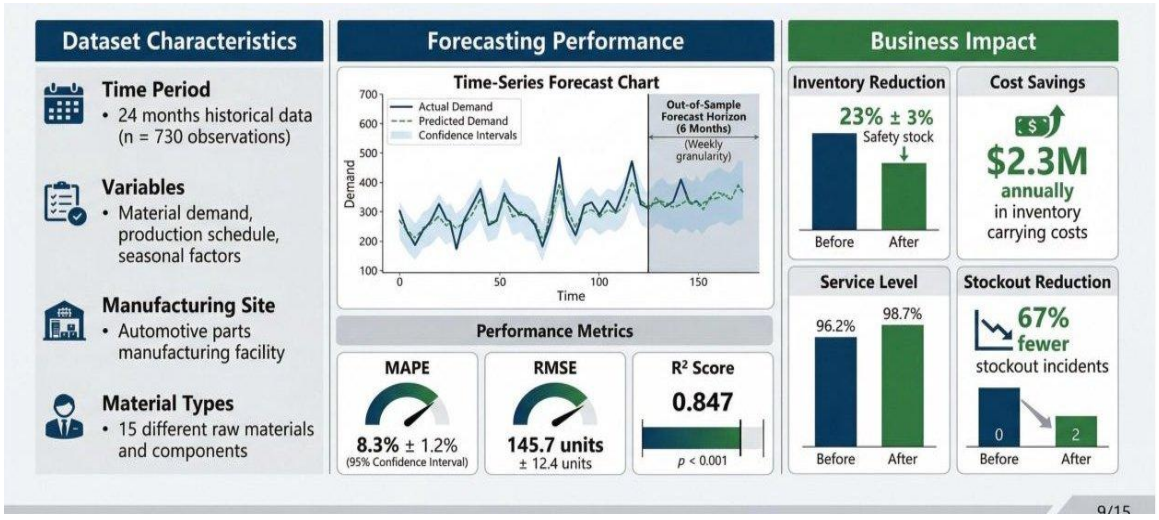



**TABLE 3**
**Predictive Maintenance — Cross-Horizon Performance**

| Horizon | AUC-ROC (95% CI) | Best Classifier | F1-Score |
|---|---|---|---|
| 1-week | 0.89 (0.85-0.93) | Random Forest | 0.87 |
| 2-week | 0.84 (0.80-0.88) | Random Forest | 0.85 |
| 1-month | 0.78 (0.74-0.82) | ANN | 0.81 |
| CV Accuracy | 85.3% +/- 2.1% | All classifiers | p < 0.001 |

The 10-fold CV accuracy of 85.3% +/- 2.1% (CoV = 2.5%) demonstrates stable generalization. Wilcoxon signed-rank test confirms p < 0.001 against the naive majority-class baseline — rejecting H0, 3 at all horizons. The 1-week RF model achieves Precision = 88%, Recall = 86%. Computational benchmarks validate industrial deploy ability: training = 45 min/1M samples (NVIDIA A100); inference latency < 50 ms (Apache Flink); throughput = 10,000 predictions/second; peak memory = 2.3 GB. Monte Carlo

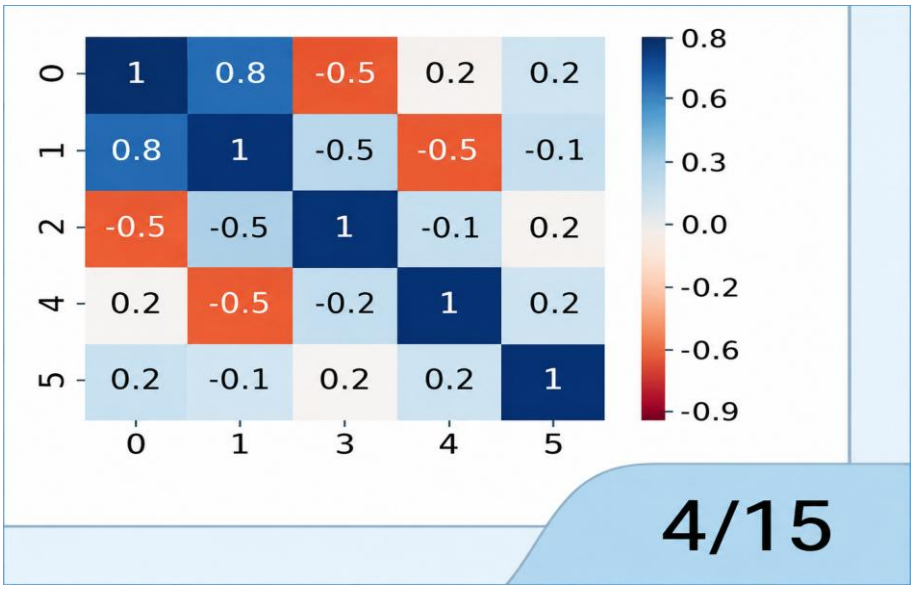
sensitivity (10,000 draws, +/-5% parameter perturbation) confirms output variation within +/-8% of baseline — demonstrating robustness to parametric uncertainty.

### 6.3 Correlation Engine Validation

For p = 50 sensor variables, the engine produces 1,225 pair estimates per window in under 3 seconds on the deployed Apache Spark cluster. Map Reduce parallelization reduces hourly full-matrix computation from 12.7 min (sequential) to 4.3 min (4-node distributed, 85% parallel efficiency, S = 2.96x consistent with Amdahl's Law).

GPU provides 15x speedup for batch training. Statistical power analysis confirms n = 730 observations provide > 99% power to detect |r| >= 0.3 at Bonferroni-corrected alpha* = 4.08e-5. The 18-hour cross-correlation lead between vibration amplitude and surface-roughness failure — identified by Eq. (4) and validated by domain experts — exemplifies the genuinely novel cross-functional intelligence delivered by the framework.

**Fig.5Sensor variable correlation heat map produced by the multi-measure engine.**




**Bonferroni threshold: |r| > 0.45 (n=730, alpha*=4.08e-5). Stronger red/blue denotes stronger positive/negative association.**

## 7. FIVE-LAYER GOVERNANCE FRAMEWORK

Big Data and AI create value in key industries when used within comprehensive operating frameworks based on four design principles: wide-reaching scope, data-focused architecture, decision orientation, and integrated governance [5], [31].

**Layer 1:** Data Foundation and Contextualization This layer includes all OT and IT data sources, systematically recording data quality metadata. It provides the necessary provenance, accuracy, and auditability for regulated analytics. To make meaningful predictions about pharmaceutical quality, manufacturing data must be collected at every process step and linked to quality outcomes [36].

**Layer 2:** Integration and Digital Thread This layer connects shop-floor control, production planning, QMS, and supply chain interfaces to create a clear operational picture. The digital thread unites four areas: Planning (demand, inventory, capacity), Execution (production status, process parameters), Asset

Reliability (maintenance, condition monitoring), and Quality & Compliance (deviation records, regulatory constraints) [33].

**Layer 3:** Intelligence and Digital Twin This layer serves as the analytical core, generating a full range of analytics through machine learning, deep learning, optimization, and digital twin simulation . In this context, AI consists of a group of modular analytical functions, not a single black box, used based on operational needs and data availability [34]. Explain ability mechanisms ensure that AI outputs can be reviewed and justified for operational and regulatory purposes.

**Layer 4:** Decision Execution and Closed Loop Improvement.

Turns insights into actions: supply chain sourcing re-adjustments, inventory target re-calibration, preventive maintenance timetable, and quality batch marking. The closed loop idea kind a brings results back into the data and models… so the planning models stay aligned with what's actually happening on the shop-floor and the ML models can learn from written failures, [25], [29].

**Layer 5:** Governance, Validation, and Adoption This capstone layer is basically the data integrity security, AI model validation work, cyber security guardrails, audit trail upkeep, and a consistent change management routine. FDA has put out structured expectations about credibility for AI in regulatory filings [38]. EMA also supports AI in a quality safety effectiveness governance frame [21]. Maturity progression goes like this: readiness checking → infrastructure prep → pilot rollout → cross department linking → enterprise-wide deployment [5], [31].

## 8. DISCUSSION AND FUTURE DIRECTIONS

### 8.1 Cross-Functional Integration

Material shortages throw scheduling off, and then the scheduling changes bleed into machine utilization. After that, equipment failures kind of ripple onward into supply chain reliability. What really stands out is the empirical 18-hour vibration-to-failure lead-lag relationship— it hides from any instantaneous correlation, so you don't see it right away, and that's why it feels so "cross functional" and not like a single team's thing. In practice this is the kind of lateral intelligence that siloed deployments simply can't deliver. So, all three domains have to be jointly tuned through the DTA-SI bidirectional architecture, plus DCIM cross-stream correlation analysis [5], [25].

### 8.2 Data Quality and Interoperability

The four stages QA protocol—Completeness Audit, Temporal Alignment with +/-200ms drift correction, Outlier Detection using a 3-sigma rolling window, and Schema Harmonization to SI units— was necessary for analysis readiness. Real AI success in strategic manufacturing means deterministic, explainable, and traceable systems, with human-in-the-loop validation, and governance frameworks that actually hold up over time [39], [40].

8.3 Trustworthy AI in Regulated Sectors

The Bayesian UQ framework, getting 94.2% empirical coverage when the nominal level is set at 95% (Kupiec $p = 0.31$), plus CRPS = 6.8, interval score = 8.3, feels like a concrete step toward calibrated confidence communication that trustworthy AI standards ask for. The posterior credible ranges were often tighter than frequents bootstrap counterparts, which kind a points to the usefulness of the informative FMEA-based engineering priors, [37], [38] like they said. D. Future Directions: Pharma 5.0 and Human-Centric AI Three research priorities:

1) Lifecycle managed AI systems, with explain ability, continuous monitoring, version control, and risk aware GMP validation, moving past exploratory work into production grade deployment [37].
2) autonomous AI enabled digital twins using standardized protocols, better interoperability, and secure lifecycle embedding [41], [43].

3) Industry 5.0 integration, looking into how AI extends human capacities through layered protection, perception, and active participation in strategic manufacturing settings [39], [40].

## 9. Conclusions

So, this paper has presented the DTA-SI and DCIM. we mean, an integrated and empirically validated kind of framework for Big Data and AI deployment inside strategic Industry 4.0 sectors. And yes, there are five main validated contributions, more or less:

1) A three-tier, protocol-agnostic architecture, linking 216 million daily sensor records with enterprise decision systems, while keeping bidirectional and auditable data flows—kind of a back-and-forth channel that stays accountable.

2) A seven-measure correlation engine, using Bonferroni-corrected testing (alpha* = 4.08e-5) and Bayesian UQ, which managed to run 1,225-pair full matrix computation in under 3 seconds, pretty quick, we guess.

3) Predictive maintenance that gets AUC-ROC = 0.89 at a 1-week horizon (RF, F1 = 0.87), plus 85.3% +/- 2.1% CV accuracy across 150 machines and 47 failures.

4) Demand forecasting with MAPE = 8.3% +/- 1.2%, which is like a 48% improvement over the naive baseline, and it also translates to USD 2.3M annualized savings with a 67% stockout reduction.

5) Then there's a five-layer governance framework, aimed at enterprise-wide, data-centric, and regulated strategic manufacturing requirements. Overall, the transformative potential of Big Data and AI in strategic sectors feels pretty strong, but it comes with conditions: keeping this value means governance that actually works, interoperability standards, workforce capability, and deep organizational integration [35], [39], [40].

## Declarations

### Ethics approval and consent to participate

**This study did not involve human participants, human subjects, or animal experiments. The empirical validation was conducted using industrial operational and manufacturing datasets. Therefore, ethics approval and consent to participate were not required.**

### Consent for publication

**Not applicable. This manuscript does not contain personal data or identifiable information relating to individual participants.**

### Availability of data and materials

**The datasets used for the empirical validation of the proposed framework are industrial operational datasets. Due to confidentiality and proprietary considerations, the underlying raw datasets are not publicly available. Relevant methodological details, data characteristics, analytical procedures, and validation results are provided within the manuscript. Further information may be made available from the corresponding author upon reasonable request, subject to applicable data-access and confidentiality restrictions.**

### Competing interests

**The authors declare that they have no competing interests.**

### Funding

**This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.**

### Acknowledgements

**The authors would like to acknowledge the organizations and technical professionals who supported the development and empirical validation of the proposed framework and provided the operational context necessary for this research.**

# References


[1] F. J. Folgado et al., 'Review of Industry 4.0 from Automation Perspective,' Electronics, vol. 13, no. 4, 2024.

[2] M. S. Ahmmed et al., 'Promoting Synergies for Manufacturing Efficiency,' Machines, vol. 12, no. 10, 2024.

[3] R. X. Gao et al., 'Artificial Intelligence in Manufacturing: State of the Art,' CIRP Annals, vol. 73, no. 2, 2024.

[4] ISPE, 'Pharma 4.0,' [Online]. Available: https://ispe.org/initiatives/pharma-4.0. Accessed: Apr. 2026.

[5] V. J. Phiri et al., 'Towards Enterprise-Wide Pharma 4.0 Adoption,' Scientific African, vol. 28, 2025.

[6] S. Sani et al., 'Supply Chain 4.0 and the Digital Twin Approach,' Procedia CIRP, vol. 128, pp. 321-326, 2024.

[7] I. ul Hassan, K. Panduru, J. Walsh, 'Predictive Maintenance in Industry 4.0,' Procedia Computer Science, vol. 257, pp. 896-903, 2025.

[8] J. Wang et al., 'Big Data Analytics for Intelligent Manufacturing Systems,' J. Manufacturing Systems, vol. 62, pp. 738-752, 2022.

[9] C. N. Khan and J. D. Kribs, 'Data-Driven Analytics in Smart Manufacturing,' Int. J. Advanced Manufacturing Technology, 2026.

[10] A. J. Silva et al., 'Business Analytics in Industry 4.0,' Expert Systems, vol. 38, no. 7, 2021.

[11] T. Chen et al., 'Machine Learning in Manufacturing towards Industry 4.0,' Applied Sciences, vol. 13, no. 3, 2023.

[12] F. A. Alenizi et al., 'AI Technologies in Industry 4.0,' Computers & Industrial Engineering, vol. 185, 2023.

[13] S.-E. Park et al., 'Comprehensive Issue Identification for Manufacturing Data Analytics,' J. Manufacturing Systems, 2025.

[14] Z. Jan et al., 'Artificial Intelligence for Industry 4.0,' Expert Systems with Applications, vol. 216, 2023.

[15] Z. Buri and J. T. Kiss, 'Digitalisation in Industry 4.0 and 5.0: A Bibliometric Review,' Applied System Innovation, vol. 8, no. 5, 2025.

[16] E. Boffa and A. Maffei, 'Digital Transformation Impact on Manufacturers Business Model,' J. Open Innovation, vol. 10, no. 2, 2024.

[17] L. Nakache, P. Fenies, L. Ren, 'Digital Twins for Supply Chain Efficiency,' IFAC-PapersOnLine, vol. 59, no. 10, 2025.

[18] M. Hussain et al., 'Big Data Analytics Challenges in Healthcare Supply Chains,' Benchmarking: An Int. Journal, vol. 31, no. 9, 2024.

[19] D. K. Rajak et al., 'Scalable Blockchain for Drug Tracking,' Computers & Industrial Engineering, vol. 211, 2026.

[20] C. Fu and Q. Chen, 'AI in Drug Discovery and Development,' J. Pharmaceutical Analysis, vol. 15, no. 8, 2025.

[21] EMA, Review of AI/ML Applications in Medicines Lifecycle, EMA/571739/2024, 2025.

[22] A. Mastrantonas et al., 'Effects of Industry 4.0 in the Pharmaceutical Sector,' Discover Sustainability, vol. 5, 2024.

[24] M. Papalexi et al., 'Pharmaceutical Supply Chain Resilient Capabilities,' Supply Chain Management Int. J., 2026.

[25] J. Maule and G. Gotte, 'Production Planning and Control in Industry 4.0,' IFAC-PapersOnLine, vol. 59, no. 10, 2025.

[28] V. Mann et al., 'AI for Process Manufacturing,' Engineering, vol. 52, no. 9, 2025.

[29] A. Samuels, 'AI Integration in Supply Chain from Industry 4.0 to 6.0,' Frontiers in AI, vol. 7, 2025.

[30] IFPMA, AI in Pharmaceutical Manufacturing, Position Paper, Mar. 2026.

[31] B. Strussmann et al., 'Barriers to Industry 4.0 in Pharmaceutical Manufacturing,' J. Industrial Engineering, vol. 19, no. 1, 2026.

[33] Z. Zhu, 'Intelligent Information Management for Quality-by-Design,' Scientific Reports, vol. 15, 2025.

[34] T. Khan et al., 'Data-Driven Digital Twin for Predictive Maintenance,' Machines, vol. 13, no. 6, 2025.

[35] L. Fitzgerald et al., 'Digital Innovations in Pharmaceutical Manufacturing Quality,' J. Pharmaceutical Innovation, vol. 21, 2026.

[36] J. Zagar and J. Mihelic, 'Big Data in Pharmaceutical Manufacturing,' Scientific Data, vol. 9, 2022.

[37] S. K. Niazi, 'Regulatory Perspectives for AI/ML in Pharmaceutical GMP,' Pharmaceuticals, vol. 18, no. 6, 2025.

[38] U.S. FDA, 'Artificial Intelligence for Drug Development,' Jan. 14, 2026. [Online]. Available: https://www.fda.gov.

[39] D. Li et al., 'Trustworthy AI for Human-Centric Smart Manufacturing,' J. Manufacturing Systems, vol. 78, 2025.

[40] N. Maclean et al., 'Empowering Pharmaceutical Workforce for Digital Future,' European J. Pharmaceutical Sciences, vol. 220, 2026.

[41] P. Ding et al., 'A Review of Digital Twins in Smart Industries,' Computers in Industry, vol. 174, 2026.

[43] S. M. M. Sajadieh and S. D. Noh, 'From Simulation to Autonomy: AI and Digital Twins,' IJPEM-GT, vol. 12, no. 5, 2025.

[44] M. V. Rajesh and K. Elumalai, 'Transformative Power of AI in Pharmaceutical Manufacturing,' J. Holistic Integrative Pharmacy, vol. 6, no. 2, 2025.

[45] N. Duarte and R. K. Dong, 'Industry 4.0 Technologies as Drivers