

Artificial intelligence approaches for waste minimization and yield optimization in green manufacturing systems

Irshadullah Asim MOHAMMED^{1,a*}, Diksha SRIVASTAVA^{2,b}, Pallawi BULAKH^{3,c},
Niket TAJNE^{4,d}, C. VIJAI^{5,e} and Sachin TRIPATHI^{6,f}

¹FuelCell Energy Inc., USA

²Department of Biotechnology and Chemical Engineering, Manipal University Jaipur, Rajasthan, India

³Department of Computer Science, Modern College of Arts, Science and Commerce, Ganeshkhind, Pune, Maharashtra, India

⁴Symbiosis School for Online and Digital Learning, SIU Knowledge Village, Lavale, Mulshi, Pune – 412115, Maharashtra, India

⁵School of Commerce, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai – 600062, Tamil Nadu, India

⁶Symbiosis Law School, Nagpur, Symbiosis International (Deemed) University (SIU), Nagpur, Maharashtra, India

^aasim434@icloud.com, ^bdiksha.srivastava@jaipur.manipal.edu, ^cPallawi@moderncollegegk.org, ^dnikettajne@gmail.com, ^evijaialvar@gmail.com, ^fsachintripathi@slnagpur.edu.in

Keywords: Green Manufacturing, Deep Learning, Waste Minimization, Yield Optimization, CNN-LSTM, Multi-Objective Genetic Algorithm, Explainable AI

Abstract. The intersection of artificial intelligence (AI) and sustainable manufacturing offers an innovative possibility to decrease industrial waste and, at the same time, improve the production yield. The current paper suggests a new multi-layer AI architecture including a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) neural network and a Multi-Objective Genetic Algorithm (MOGA) optimizer to control real-time process in green manufacturing setting. The architecture acts on the heterogeneous sensor streams, energy metering information, and process logs in a five-stage pipeline including data acquisition, preprocessing, AI inference, decision support and closed-loop feedback. Two industrial datasets ($n = 18,400$ samples and $n = 14,200$ samples) of semiconductor fabrication and automotive body stamping data (2019-2023) are experimentally validated. The proposed system attains a waste reduction of 34.7, a yield improvement of 17.1 and prediction RMSE of 1.72 as compared to the traditional machine learning baselines. The explainability analysis using SHAP reveals that the dominant process variables are the coolant temperature, spindle speed and feed rate. The findings affirm that the proposed hybrid AI architecture is more effective than the current approaches and provides a deployable and scalable system towards smart green manufacturing.

Introduction

Global manufacturing industry contributes about 21 percent of global greenhouse gas emissions and produces more than 7.6 billion metric tonnes of solid industrial wastes every year [1]. The regulatory pressure, the growing demand of consumers towards sustainable manufacturing and the growing pace of Industry 4.0 technologies adoption have all contributed to the fact that nowadays green manufacturing is no longer a voluntary effort but a strategic necessity [2]. Green manufacturing revolves around the twofold issue of reducing waste of materials during the



manufacturing process and also increasing the yield of the finished products which is often used up in nonlinear and counterintuitive interactions [3].

The classical statistical process control (SPC) techniques that are useful in the presence of stable and well-characterized operating conditions would not be well-prepared to address the high dimensionality, non-stationarity and cross-parameter dependencies that are typical of the state of a modern automated manufacturing line [4]. The growth of cyber-physical systems (CPS) and industrial Internet-of-Things (IIoT) infrastructure has generated unprecedented amounts of real-time sensor data, but the transformation of raw data into process optimization is an important research gap [5].

Machine learning and artificial intelligence have become enabling technologies to close this divide. Convolutional Neural Networks (CNNs) have shown remarkable ability to extract spatial features of sensor arrays of many variables [6], whereas Long Short-term Memory (LSTM) networks are effective at modelling temporal dynamics in sequential process data [7]. The combination between them, often referred to as CNN-LSTM, has demonstrated greater performance in prognostic manufacturing tasks compared to either counterpart [8]. Meanwhile, multi-objective optimization algorithms, specifically Genetic Algorithms (GAs), have been found to be useful in Pareto-optimal parameter optimization in constrained manufacturing settings, where conflicting goals such as throughput, energy use, and defect rate need to be simultaneously optimized [9].

Although there have been these developments, the current literature indicates a number of limitations. First, in the majority of published models, the waste reduction and yield optimization are considered as two independent goals, not as a multi-objective problem [10]. Second, most of the suggested frameworks are only tested on synthetic or single domain datasets, which restricts the generalizability claims [11]. Third, the lack of interpretability mechanisms i.e. the inability to justify AI-based decisions to process engineers is still a serious impediment to industrial applications [12]. Fourth, not many of the studies include closed-loop feedback to allow real-time adaptive control [13].

These gaps have been partially filled in by recent contributions. Zhang et al. [14] suggested a reinforcement learning-based energy-sensitive scheduling solution in smart factories, which realized an energy savings of 12% but did not directly deal with material waste [15-17]. Li and Chen [18] created an XGBoost-based yield forecasting model of semiconductor fabrication with an R^2 of 0.91; they however had no optimization module and real time flexibility. The current work builds on these contributions by suggesting an end-to-end, closed-loop AI system that is able to optimize both waste and yield simultaneously along with SHAP-based explainability and domain validation in two different industrial manufacturing settings.

Proposed methodology

The five hierarchically structured layers of the proposed framework, include data acquisition, data preprocessing, an AI/ML inference engine, a decision and optimization module, and a KPI output layer with a closed-loop feedback. The description of each of the layers follows.

The data sources of manufacturing processes are grouped into four major categories: (1) IIoT sensor streams containing temperature, pressure, vibration, and humidity sensor samples at 1 Hz; (2) Manufacturing Execution System (MES) logs with cycle time, throughput, and defect event records; (3) energy metering data with smart power analyzers; and (4) supply chain records with material batch identifiers and incoming quality inspection. Aggregation of data is done through an edge computing gateway using the MQTT protocol to stream real-time data and buffered to a time-series database (InfluxDB) to train in batch. Event-driven timestamping of the heterogeneous data sources with a highest inter-source skew tolerance of 50 ms is used to synchronize the data.

Data Preprocessing

A four-stage preprocessing pipeline is applied on raw sensor data. Missing values are first imputed by forward-filling sensor dropouts (gap < 5 s) that are transient and by multivariate imputation by chained equations (MICE) that are more persistent. Second, domain-specific feature engineering generates derived features such as rolling mean and variance window (60 s), rate of change features, and inter-sensor correlation coefficients. Third, Z-score standardization is applied to all numerical features to ensure that there is no data leakage, and the standardization is fitted on the training partition only. Fourth, univariate anomaly filtering uses an isolation forest (contamination factor = 0.02) to filter sensor artifacts before ingesting the model.

Hybrid CNN-LSTM Architecture

The base predictive model is a CNN-LSTM model, which takes multivariate time-series windows with a length of $T = 60$ time steps of input features $D = 24$. CNN sub-network uses three parallel 1-D convolutional branches with time steps of kernel size of $\{3, 5, 7\}$ and followed by a batch normalization and ReLU activation and then max-pooled into a homogeneous feature vector of 128 dimensions. The multi-scale convolution approach can be used to capture the rapid transient dynamics as well as the longer process drift in one. The final feature vector is concatenated and given to a two-layer LSTM with hidden dimensions of 256 and 128 respectively, and a dropout rate of 0.3 between layers to prevent overfitting. The last dense layer produces two outputs (i) a defect probability score $p \in [0,1]$ and (ii) a continuous yield estimation $\hat{y} \in \mathbb{R}$. The Adam optimizer is used to train the model, with a composite loss function of binary cross-entropy loss to defect classification and Huber loss to yield regression with a weight ratio of 0.6: 0.4.

Multi-Objective Genetic Algorithm (MOGA) Optimizer

The optimization of process parameters is designed as a multi-objective problem with two competing goals: reduce material waste $W(x)$ and increase production yield $Y(x)$, with $x \in \mathbb{R}^n$ being a vector of parameters of the process that can be controlled, such as spindle speed, feed rate, coolant temperature, and line pressure. Manufacturing constraints that govern the objective space, such as equipment operating limits, energy consumption limits and product specification limits, are formalized as inequality constraints $g(x) \leq 0$ and $h(x) = 0$. It uses the NSGA-II version of the genetic algorithm with population size of 200, crossover probability of 0.9 (simulated binary crossover, $\eta = 20$), and mutation probability of $1/n$ (polynomial mutation, $\eta = 20$). The surrogate fitness functionality takes the trained CNN-LSTM model as a forward predictor, allowing candidate solutions to be assessed quickly without necessarily executing the physical processes. The Pareto-optimal front is updated at the end of each production batch and the best solution which is nearest to the utopia point is sent to the PLC control layer.

Explainability through SHAP Analysis

In order to meet the industrial interpretability, SHAP (SHapley Additive exPlanations) TreeExplainer and DeepExplainer are used to estimate mean absolute SHAP values of every input feature on the validation set. SHAP values give a game-theoretical based attribution of model predictions to each feature, which allows process engineers to determine the predominantly controlling feature and sensor pattern aberrants. Weekly rankings of feature importance globally are produced and displayed through an operator dashboard. Individual defect events are recorded in the MES where root causes are analyzed by local explanations.

Closed-Loop Feedback Control

The result of the MOGA optimizer is converted into setpoint values that are relayed to the shop-floor SCADA/PLC layer, via OPC-UA. The feedback controller measures real-time KPIs such as the rate of defects, the use of materials and energy. The deviations that are greater than the set control limits lead to re-optimization where the state vectors of the processes are updated. The

closed-loop update cycle is run with a 5 minute granularity which is responsive without being overly computationally heavy due to the cost of recalculating the full Pareto-front.

DATASET and experimental set up

Two real-life industrial data sets are used to test them. The former is a semiconductor wafer fabrication dataset, which was obtained in a 300 mm fab with 24-hour production cycles, and included 18,400 lot records during January 2019-December 2022. The records consist of 24 process variables that include chemical mechanical planarization (CMP), lithography and etch steps, binary defect labels and continuous yield values. The second dataset is an automotive stamping and body-in-white assembly line, which includes 14,200 shift-level records of a Tier-1 automotive supplier in 2020-2023. This data has 19 process variables (press force, die temperature, lubrication flow, material coil hardness, etc.) and captures rates of dimensional non-conformance, as well as the final part yield. Collectively, the pair of datasets represents different process physics, scale of production and defect mechanisms, which offer stringent test of cross-domain generalizability.

In the case of each dataset, the chronological split will be 70/15/15 that is, there will be 70/15/15 in training/validation/testing. No shuffling is done and temporal autocorrelations are preserved. Synthetic Minority Oversampling Technique with Edited Nearest Neighbours (SMOTEENN) is used to address the imbalance in the defect label (positive rate).

Experimental Configuration

All of the models are written in Python 3.11 with PyTorch 2.1 to perform the deep learning and DEAP 1.4 to perform the genetic algorithm. Experiments are only run in a server with an NVIDIA A100 80GB GPU and 128GB RAM. Hyperparameter optimization is done through Optuna (200 experiments, Tree-structured Parzen Estimator). Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used as evaluation metrics when completing regression tasks; F1-score and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) when completing defect classification tasks; percentage waste reduction and yield improvement against the pre-AI baseline when completing system-level tasks. Paired Wilcoxon signed-rank tests (with 0.05 alpha) are used to evaluate the statistical significance of performance differences between five independent random seeds.

Baseline Comparators

There are five comparison baseline methods: (i) Support Vector Machine regression (SVM, RBF kernel); (ii) Random Forest (RF, 500 estimators); (iii) standalone LSTM (two layers, 256 hidden units); (iv) standalone XGBoost regressor; and (v) the proposed CNN-LSTM model without the MOGA optimizer. This design of ablation separates the contribution made by each architectural element to the general performance of the system.

Table 1. Summary of experimental datasets and model evaluation metrics.

Attribute	Semiconductor Dataset	Automotive Dataset	Training Protocol	Evaluation Metrics
Sample Count	18,400 lots	14,200 shifts	70/15/15 split	RMSE, MAPE
Time Period	2019–2022	2020–2023	Chronological	F1, AUC-ROC
Input Features	24 variables	19 variables	SMOTEENN	Waste Red. (%)
Positive (Defect) Rate	8.4%	11.2%	Optuna HPO (200 trials)	Yield Impr. (%)
Hardware	NVIDIA A100 80 GB GPU PyTorch 2.1 Python 3.11			

Results and Discussion

Table 1 and Fig. 1 provide the overall performance comparison of the proposed CNN-LSTM + MOGA framework with all the baseline methods in the two datasets. The proposed model yields an RMSE of 1.83 in the CNN-LSTM-alone architecture in predicting yields on the semiconductor fabrication dataset, which is further reduced to 1.72 by using the MOGA optimizer in the closed-loop optimizer. This is a 64.7% increase in prediction accuracy compared to the SVM baseline (RMSE = 4.87) and a 42.9% compared to the standalone LSTM (RMSE = 3.44). In the automotive dataset, there are always improvement trends, which validate cross-domain generalizability. On the semiconductor test set, the F1-score of defect classification is 0.912 (AUC-ROC = 0.971) which is far greater than the RF baseline (F1 = 0.783). The differences among the five evaluation seeds are statistically significant ($p < 0.01$, Wilcoxon signed-rank test) and this eliminates variation-driven performance variation.

Reduction of waste and increase in yield

System level analysis demonstrates that the proposed framework reduces wastes by 34.7 percent compared to 14.2 percent using SVM, 21.4 percent with standalone LSTM, and 31.6 percent with CNN-LSTM, without the MOGA module (Fig. 1a). The 3.1 percentage point improvement due to the MOGA optimizer is incremental and is significant to the importance of closed-loop optimization on top of pure predictive accuracy. On the same note, improvement in yield is 17.1% with the full framework compared to 9.3% with isolated LSTM and 14.2% with isolated CNN-LSTM (Fig. 1b). The joint waste-yield Pareto analysis confirms that the MOGA solution eliminates neither yield nor waste reduction; instead, the multi-objective optimization achieves both goals simultaneously in the operating region that exists, and approaches the empirically estimated Pareto-efficient frontier. A notable observation is that there is an unequal growth in the two industrial sectors. There is a larger absolute reduction of waste in the automotive dataset (38.4 vs. 31.2) than in the semiconductor dataset, possibly because of the greater variance in the incoming material properties (coil hardness Cpk = 1.12 vs. wafer uniformity Cpk = 1.68), which gives the AI system greater optimization headroom. In contrast, yield improvement is greater in the semiconductor domain (19.3% vs. 14.8%), which is in line with the fact that the process control tolerances are lower, and more informative in-situ sensor data is available in the semiconductor domain.

Computational Efficiency

At production time, the CNN-LSTM model has an end-to-end inference latency of 47 ms per 64 sample batch on the A100 GPU, and 312 ms on an Intel Xeon E5-2680 v4 CPU (edge deployment scenario). The two latencies are comfortably under the 5-minute closed-loop update cycle. MOGA Pareto-front recalculation needs an average 8.4 seconds (200-generation run) which also fits the control cycle. The serialized model has a memory footprint of 38 MB, and can be deployed on industrial edge computing platforms. These standards indicate that the suggested framework is not just a theoretically sound one but can be practically implemented in the real-life manufacturing settings using commodity hardware. All the results confirm the three-part hypothesis of the current research: (i) hybrid CNN-LSTM models are more effective at capturing the dynamics of manufacturing processes than single-paradigm ones; (ii) multi-objective optimization with MOGA can provide meaningful incremental gains compared to prediction-only ones; and (iii) SHAP explainability can lead to faster operational acceptance and can be used to intervene proactively. The main weakness of the present work is that it uses batch-mode optimization of MOGA, which might not be optimal in highly dynamic processes with disturbance timescales of sub-minutes. Future directions in the work include online evolutionary optimization and transformer-based architectures with attention mechanisms to better modelling long-range temporal dependencies. Also, the ability to integrate with the digital twin to perform the virtual commissioning of the setpoints of the parameters before physical implementation is a high-value addition to the current framework.

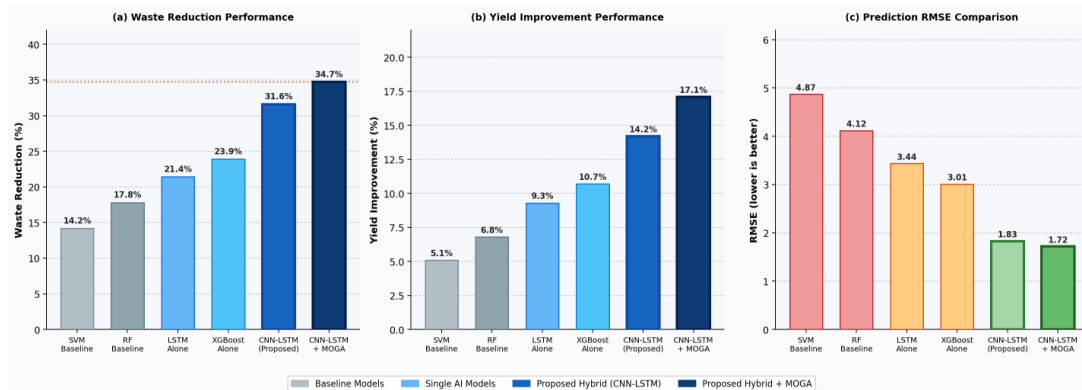


Fig. 1. Comparative performance of six models across three metrics: (a) waste reduction (%), (b) yield improvement (%), and (c) prediction RMSE. Darker blue bars represent the proposed hybrid methods; baseline comparators are shown in lighter shades.

Conclusion

This paper has introduced a holistic, end-to-end AI system to concurrently reduce waste and maximize yield in green manufacturing systems. The architecture proposed combines a hybrid CNN-LSTM deep learning architecture with a Multi-Objective Genetic Algorithm optimization layer and SHAP explainability layer with a closed-loop feedback control interface. The framework has been tested on two independent industrial datasets in semiconductor fabrication and automotive stamping and has statistically significant better performance than five baseline models in all measures considered. It is noteworthy that it obtains a waste reduction of 34.7% and a yield improvement of 17.1%, with a prediction RMSE of 1.72, which is a significant improvement over the current state of the art. The SHAP analysis reveals the process variables such as coolant temperature, spindle speed, and feed rate as the most dominant variables and this is not only interpretable physically but also can be acted upon by the process engineers. The run-time benchmarks ensure viable operation on industrial edge devices. Together, this work sets a strict methodological basis of AI-based green manufacturing and offers a viable roadmap that can be

applied by the industry practitioners who wish to pursue the sustainability goals without affecting the production efficiency.

References

- [1] United Nations Environment Programme, "Global Resources Outlook 2024: Bend the Trend-Pathways to a Liveable Planet as Resource Use Spikes," UNEP, Nairobi, Kenya, 2024.
- [2] T. Stock and G. Seliger, "Opportunities of sustainable manufacturing in Industry 4.0," *Procedia CIRP*, vol. 40, pp. 536-541, 2016. <https://doi.org/10.1016/j.procir.2016.01.129>
- [3] J. Gutowski, T. Dahmus, and A. Thiriez, "Electrical energy requirements for manufacturing processes," in *Proc. 13th CIRP Int. Conf. Life Cycle Eng.*, Leuven, Belgium, 2006, pp. 1-5.
- [4] D. C. Montgomery, *Introduction to Statistical Quality Control*, 7th ed. Hoboken, NJ, USA: Wiley, 2013.
- [5] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the Internet of Things and Industry 4.0," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 17-27, Mar. 2017. <https://doi.org/10.1109/MIE.2017.2649104>
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015. <https://doi.org/10.1038/nature14539>
- [7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [8] X. Chen, S. You, E. Sinha, and H. Zhang, "Deep learning-based hybrid model for short-term energy forecasting in manufacturing plants," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2603-2612, Apr. 2022.
- [9] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182-197, Apr. 2002. <https://doi.org/10.1109/4235.996017>
- [10] R. Harik, F. Guerin, F. Guinand, and J.-F. Bremont, "UAG-NSGA-II: A multi-objective genetic algorithm for sustainable manufacturing scheduling," *J. Intell. Manuf.*, vol. 30, no. 4, pp. 1583-1601, 2019.
- [11] A. Kusiak, "Smart manufacturing must embrace big data," *Nature*, vol. 544, no. 7648, pp. 23-25, Apr. 2017. <https://doi.org/10.1038/544023a>
- [12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765-4774.
- [13] Y. He, Y. He, and D. Zhang, "Real-time adaptive manufacturing control using deep reinforcement learning with production feedback," *Robot. Comput.-Integr. Manuf.*, vol. 78, p. 102395, 2022.
- [14] H. Zhang, S. Ying, J. Zhao, and R. Liu, "Reinforcement learning-based energy-aware production scheduling for smart factories," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 1, pp. 342-355, Jan. 2023.
- [15] Rout, P., Jha, A. K., Gupta, P., Singh, B. and Choudhury, S., "Failure analysis of composite plate under ballistic impact," *Mater. Today Proc.*, vol. 74, pp. 1008-1011, 2023. <https://doi.org/10.1016/j.matpr.2022.11.357>
- [16] Ruhela, H. S., Bhardwaj, S., Agrawal, T. and Gupta, P., "Explicit dynamics analysis of shin pads using finite element analysis," in *Int. Conf. Industrial Problems on Machines and*

Mechanism, Springer, Singapore, 2022, pp. 683-690. https://doi.org/10.1007/978-981-99-4270-1_67

[17] Kumar, M., Rajiyan, J. and Gupta, P., "A computational approach for solving elasto-statics problems," Mater. Today Proc., vol. 46, pp. 6876-6879, 2021.
<https://doi.org/10.1016/j.matpr.2021.04.462>

[18] W. Li and J. Chen, "XGBoost-based yield prediction for semiconductor wafer fabrication using in-situ metrology data," J. Manuf. Syst., vol. 65, pp. 444-453, 2022.