

AI-Driven Optimisation of Printing Production Using Neural Networks within a FAIR Data Framework

Chris Trochoutsos^{1,2,3}, Gerasimos Vonitsanos^{1,2}, Marios Tsigonias^{1,2}

1. University of West Attica, Hellenic Graphic-Media Research Lab – GRAPHMEDLAB, Greece

2. Hellenic Open University, School of Applied Arts and Sustainable Design, Greece

3. Pressious Arvanitidis, printing company, Greece

E-mail: chtrox@pressious.com, gvonitsanos@uniwa.gr, mtsigonias@uniwa.gr

Short abstract

This paper presents a neural network-based framework for predicting ink consumption in offset printing environments, leveraging structured FAIR data pipelines within the CyclOps framework. The proposed approach integrates heterogeneous production data, including machine parameters, environmental conditions, and material characteristics, into a unified data lifecycle to support AI-driven modelling. Both feedforward and recurrent neural network architectures are employed to capture static relationships and temporal dependencies within the printing process. The models are trained on multivariate industrial data to predict key sustainability indicators, especially ink consumption, and are evaluated using standard regression metrics. Results indicate improved prediction performance, particularly when temporal dynamics are considered, while the CyclOps-based data lifecycle supports structured, traceable, and interoperable data management for more efficient and sustainable printing production.

Keywords: Neural Networks, Offset Printing, FAIR Data Management, Sustainable Manufacturing, CyclOps

1. Introduction and background

Offset printing remains a fundamental process within the graphic arts and packaging industry, valued for its high quality and productivity. Nevertheless, achieving consistent output remains a complex challenge, as it requires precise control over multiple interdependent variables, including ink–water balance and substrate properties (Pudas et al., 2004). Despite the process's technological maturity, production is still largely guided by operator-driven adjustments, which can lead to material waste and reduced operational efficiency.

In the context of Industry 4.0, modern printing environments generate substantial volumes of data from machine sensors and enterprise systems. However, these data are often fragmented and insufficiently structured, limiting their effective utilisation in advanced analytical applications. The adoption of FAIR data principles, ensuring that data are Findable, Accessible, Interoperable, and Reusable, provides a structured foundation for overcoming these limitations (Wilkinson et al., 2016). This approach is closely aligned with the development of European data spaces, which aim to facilitate secure data sharing and foster collaborative innovation (European Commission, 2020; 2023).

Conventional statistical methods, although widely used for process monitoring, are limited in their ability to capture the nonlinear and temporal dependencies inherent in printing processes. In contrast, neural network architectures offer enhanced capability in modelling complex patterns in industrial time-series data (DiPietro & Hager, 2020). Recent studies have further demonstrated the effectiveness of deep learning techniques in predictive modelling within data-intensive environments (Emmert-Streib et al., 2020).

This study addresses the gap between heterogeneous data generation and informed decision-making in printing production. By leveraging the CyclOps framework (Caballero et al., 2026), an automated data pipeline is established to transform production data into FAIR-compliant assets. This infrastructure enables the deployment of neural network models to predict and optimise ink consumption, contributing to more efficient and sustainable printing operations.

2. Materials and methods

The proposed framework is designed to support the prediction of resource consumption and production waste in offset printing environments by integrating neural network models into a structured data lifecycle. The methodology combines heterogeneous data acquisition, preprocessing, feature engineering, and machine learning modelling within a FAIR-compliant data pipeline.

2.1 Data Sources and Acquisition

The dataset used in this study represents a typical offset printing production environment, incorporating process-related and environmental variables that influence resource consumption. It consists of multi-source industrial data collected from printing machinery, environmental sensors, and production systems. All variables are time-stamped and synchronised, forming a multivariate dataset suitable for both static and temporal modelling. The dataset includes approximately 10,000–15,000 observations collected across multiple production runs under varying conditions. Each observation reflects a specific production state, capturing machine, environmental, and material parameters at regular intervals. The collected variables are categorised by their functional role in the printing process, as presented in Table 1.

Table 1: Description of variables

Category	Parameter	Specification
Machine	Printing press	Sheetfed offset press (industrial configuration)
Machine	Printing speed	8,000–16,000 sheets/hour
Machine	Ink density	Adjustable per printing unit
Machine	Dampening system	Variable water supply (0–100%)
Environment	Temperature	20–25 °C (controlled industrial conditions)
Environment	Relative humidity	45–60%
Environment	Absolute humidity	Derived from temperature and RH
Material	Paper	Coated and uncoated substrates
Material	Paper grammage	80–200 g/m ²
Material	Ink	Conventional offset inks
Production	Print coverage	10–90%
Production	Job type	Variable production scenarios
Data	Sensors	Machine logs and environmental sensors
Data	Data format	Time-series multivariate dataset

It should be noted that ink rheology variables, including tack, viscosity, ink temperature, and roller cooling conditions, are not included in the current dataset. Although these parameters are known to affect ink transfer in conventional offset printing, the inks used in the examined production environment are treated as standardised materials under controlled production conditions. Moreover, the available data describe print job characteristics and production settings, such as quantity, format, printed colours, paper properties, and machine configuration, rather than

detailed press-side ink measurements. Consequently, no systematic variation in rheological ink properties is available for model training. These factors are therefore considered operationally constant within the scope of this study and are not modelled as independent input variables.

Print coverage in this study was obtained during the pre-press stage as Total Area Coverage (TAC) derived from PDF analysis of the digital print file. This TAC value is a nominal estimate of the expected ink coverage before production. It is based on the planned print content, rather than online ink-key feedback or real-time press-side measurements. The predictive target corresponds to total ink consumption per print job, aggregated across all printed colours. Separate ink consumption values for individual CMYK separations are not predicted in the current implementation. Consequently, the proposed model is intended for overall ink consumption prediction and resource-efficiency assessment, rather than colour-separation-level ink optimisation.

The dataset is structured to reflect realistic industrial conditions, where multiple factors simultaneously influence production outcomes. By integrating machine, environmental, material, and production variables, the dataset enables the modelling of complex nonlinear relationships between process conditions and sustainability indicators. Data acquisition is conducted through embedded machine sensors, environmental monitoring systems, and digital production logs within the printing environment. This integrated data collection approach supports the development of predictive models that estimate resource consumption and identify inefficiencies in offset printing processes.

2.2 Data Preprocessing and Integration

Following data acquisition, a structured preprocessing pipeline (Kumar et al., 2023) was applied to ensure data quality, consistency, and suitability for modelling. Initially, all data streams were time-synchronised using their timestamps, enabling alignment of machine, environmental, and production variables. This step was essential for capturing interactions among heterogeneous data sources in the printing process. Subsequently, missing and inconsistent values were addressed through appropriate filtering and imputation techniques. Noise reduction was performed to minimise the impact of sensor variability. Redundant or irrelevant variables were removed to improve computational efficiency and reduce model complexity.

The processed data were then integrated into a unified multivariate dataset, enabling simultaneous analysis of multiple process factors. This integration facilitated the modelling of interdependencies between variables and supported downstream machine learning tasks.

2.3 Neural Network Modelling

The modelling approach was based on neural networks to capture nonlinear relationships between process variables and production outcomes (Goodfellow et al., 2016). Two architectures were employed to address both static and temporal aspects of the data. A Multilayer Perceptron (MLP) (Emmert-Streib et al., 2020) with two hidden layers (64 and 32 neurons, ReLU activation) was used to model static relationships, while a Long Short-Term Memory (LSTM) network (DiPietro & Hager, 2020), consisting of a single layer with 50 units followed by a dense output layer, was implemented to capture temporal dependencies.

The models were trained using supervised learning on historical production data. Input features included machine, environmental, material, and production-related variables, while the primary output corresponded to total ink consumption. To avoid data leakage, the dataset was split chronologically at the production-run level. The earliest 70% of production runs were used for training, the subsequent 15% were used for validation and early stopping, and the most recent

15% were reserved as an independent held-out test set. Data were normalised using min-max scaling, with scaling parameters estimated from the training set only and subsequently applied to the validation and test sets. Training was performed using the Adam optimiser and Mean Squared Error (MSE) loss over a maximum of 30 epochs with a batch size of 32, and early stopping was applied based on validation loss to reduce overfitting.

The models were implemented and executed within the Cyclops framework, enabling integration with the data pipeline and supporting efficient training and deployment. Model performance was evaluated using standard regression metrics, including MAE, RMSE, and R^2 (Chicco et al., 2021).

2.4 Integration within the Cyclops Framework

The proposed methodology was implemented within the Cyclops framework to address the challenges of managing and utilising heterogeneous industrial data in offset printing environments, where data from machines, production systems, and job specifications are often fragmented. Cyclops was employed as enabling infrastructure to support the structured integration and management of these data streams in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable). This supports the consistent organisation of production data and their preparation for neural network modelling, where data quality, traceability, and interoperability are critical. In addition, the framework supports a unified data lifecycle, including data ingestion, preprocessing, and storage within a Manufacturing Data Space-oriented environment, allowing the integration of production, machine-related, and material-related data into a coherent dataset relevant to ink consumption (Caballero et al., 2026).

To make the FAIR data management approach more explicit, the dataset is accompanied by structured metadata describing the source, meaning, format, and units of the production variables used in the modelling workflow. These variables include job identifiers, delivery time, quantity, quality, printed colours, press sheets, dimensions, job type, number of sides, paper weight, varnish information, time-cost information, and machine-line selection. The Job ID from the production management system is used as an internal traceability identifier, allowing each production record to be linked to its corresponding print job throughout data ingestion, preprocessing, model training, and evaluation.

In the current implementation, semantic consistency is supported through a controlled data dictionary defining variable names, descriptions, units, and allowed values where applicable. The dataset is managed within the Cyclops data lifecycle using internal versioning and traceability mechanisms rather than through a public DOI-based repository. Future work will extend this approach toward formal ontology alignment and external repository publication, where data-sharing restrictions allow.

The integration of neural network models into Cyclops supports scalable model training, validation, and future deployment across the structured data lifecycle. In this study, Cyclops serves as supporting infrastructure, facilitating the practical application of data-driven modelling by ensuring that the required data are consistently structured, traceable, and available for analysis.

3. Results and discussion

The evaluation primarily focuses on predicting ink consumption, the study's main objective. The performance of the proposed neural network models was evaluated using standard regression metrics, focusing on their ability to predict key sustainability indicators, including ink consumption, paper waste, and process efficiency. As summarised in Table 2, both models demonstrated predictive capability across the examined indicators, with R^2 values of 0.78 for the

MLP and 0.85 for the LSTM, while MAE values remained below 1.0 across all targets. This comparison is limited to the two tested neural network architectures and does not constitute a full benchmark against conventional regression models. Although min-max normalisation was applied during model training, the reported MAE and RMSE values were computed after inverse transformation of the predicted and actual values to the original ink-consumption scale. Therefore, the reported errors are expressed in kg of total ink consumption per print job, rather than in normalised units. Lastly, MAE (% of mean) expresses the MAE as a percentage of the mean observed ink consumption in the test set.

Table 2: Model Performance Comparison

Model	MAE (kg/job)	MAE (% of Mean)	RMSE (kg/job)	R ²
MLP	0.82 ± 0.12	13.6%	1.05 ± 0.16	0.78
LSTM	0.65 ± 0.09	10.8%	0.89 ± 0.11	0.85

This performance difference can be attributed to the LSTM architecture's ability to capture temporal dependencies inherent in printing production processes, in which process conditions evolve. The lower MAE values indicate reduced average deviations between predicted and actual values. In contrast, the lower RMSE reflects a decrease in larger prediction errors, which are particularly critical in industrial environments.

These findings are further supported by Figure 1, which illustrates the relationship between predicted and actual ink consumption values for the MLP and LSTM models. The values on both axes are presented on a min-max normalised scale for visualisation and model-comparison purposes, enabling consistent comparison of prediction patterns across varying production conditions. Both models capture the overall relationship between the input variables and ink consumption. However, the MLP model exhibits greater scatter around the ideal prediction line ($y = x$), indicating larger deviations between predicted and actual values. In contrast, the LSTM model shows closer alignment with the diagonal, suggesting improved agreement between predicted and observed ink consumption.

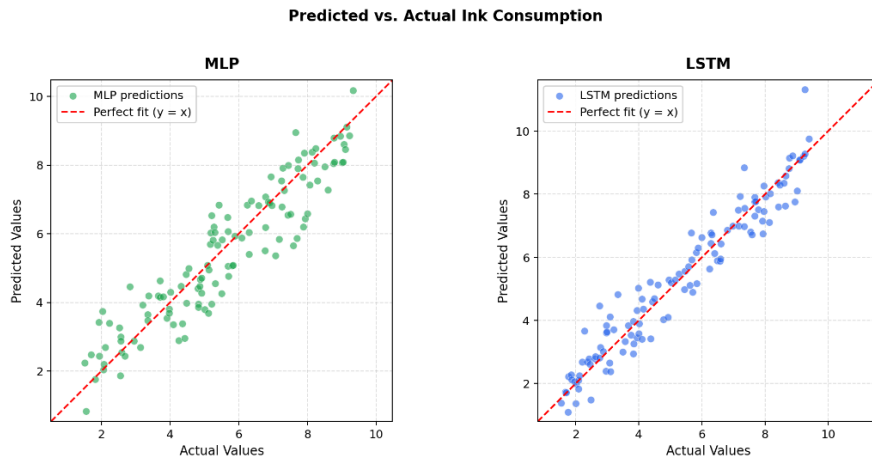


Figure 1: Predicted vs actual ink consumption for MLP and LSTM models

The enhanced performance of the LSTM model is also evident in Figure 2, which presents a comparative evaluation based on MAE and RMSE. The consistently lower error values achieved

by the LSTM model confirm its improved robustness and predictive capability, particularly in handling dynamic variations in production conditions.

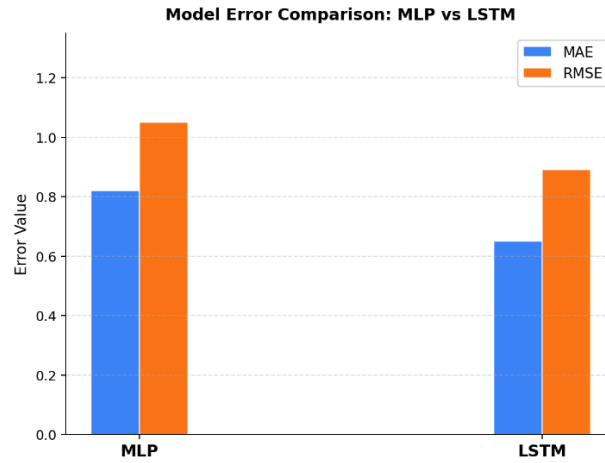


Figure 2: Comparison of MAE and RMSE values for the MLP and LSTM models

The results extend beyond predictive accuracy by indicating model-detected associations between process variables and ink consumption patterns. Variations in environmental and operational parameters, such as humidity and printing speed, were associated with changes in predicted ink consumption. However, these relationships should be interpreted as predictive associations rather than causal effects, since the neural network models used in this study do not provide causal explanations on their own. Accordingly, the findings support data-driven monitoring and exploratory analysis. At the same time, causal interpretation and targeted process optimisation require additional explainability analyses, such as permutation importance, SHAP values, or partial dependence methods. The integration of the models within the CycloOps framework further enables structured monitoring and future extension toward explainable and adaptive optimisation workflows.

4. Conclusions

This study investigates the application of data-driven modelling in offset printing processes, with particular emphasis on predicting and controlling ink consumption, a critical determinant of production efficiency and environmental performance. The findings emphasise the influence of temporal variability on ink consumption patterns, indicating that fluctuations in process conditions—such as environmental changes and operational adjustments—have a cumulative effect on ink usage. This highlights the necessity of accounting for the dynamic behaviour of printing systems to achieve consistent ink transfer and minimise material waste. From an operational standpoint, the results provide evidence that specific configurations of process parameters and environmental conditions are associated with variations in ink consumption. This insight supports a more systematic approach to process control, enabling targeted adjustments to improve ink efficiency and reduce unnecessary consumption. Furthermore, the incorporation of predictive modelling within the CycloOps framework facilitates a transition towards proactive management of ink usage. By supporting structured data integration, continuous monitoring, and scalable analytics, the framework enables earlier identification of deviations and more informed intervention strategies, thereby enhancing process stability and resource utilisation.

Future work will focus on validating the proposed LSTM-based approach in day-to-day production planning and quantifying potential ink and monetary savings using real production and ink cost data. Further extensions will include incorporating ink rheology variables, per-

separation CMYK prediction, conventional baseline models, repeated time-based or leave-one-run-out validation, and explainability analyses such as permutation importance, SHAP values, and partial dependence methods. The FAIR data management strategy will also be extended through formal ontology alignment and external repository publication, where data-sharing restrictions allow.

Acknowledgements

This research initiative is supported by the European Union's Horizon Europe Framework Programme for Research and Innovation, under the CycOps project (Grant Agreement Project 101135513). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the CycOps consortium and the European Union, which cannot be held responsible for them.

References

- Caballero, M., Queralt, A., Calvanese, D., Oikonomidis, A., Muonagor, C., Abella, A., ... and Ordóñez, D. 2026. CycOps: Leveraging Semantic Technologies for AI and Data Life Cycle Management and Governance. In Springer Nature Switzerland, *Artificial Intelligence, Data and Robotics: Foundations, Transformations and Future Directions Cham*: pp. 403-432.
- Chicco, D., Warrens, M. J. and Jurman, G. 2021. The coefficient of determination, R-squared, is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623.
- DiPietro, R. and Hager, G. D. 2020. Deep learning: RNNs and LSTM. In *Handbook of medical image computing and computer assisted intervention Academic Press* pp. 503-519.
- Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. 2016. Deep learning, *The MIT press* 1(2) pp. 1-800. Cambridge:.
- Kipphan, H. 2001. Handbook of print media: technologies and production methods (Vol. 1207). Berlin: Springer.
- Kumar, A., Kumar, P., Mittal, R. K. and Singh, H. 2023. Preprocessing and postprocessing in additive manufacturing. In *Advances in Additive Manufacturing*, Elsevier, pp. 141-165.
- Kumar, S. and Saini, M. 2016. Study of Conventional Offset Printing with Digital Printing. *International Journal of Science, Engineering and Computer Technology*, 6(2), 110-112.
- Pudas, M., Hagberg, J. and Leppävuori, S. 2004. Printing parameters and ink components affecting ultra-fine-line gravure-offset printing for electronics applications. *Journal of the European Ceramic Society*, 24(10-11), 2943-2950.
- Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S. and Dehmer, M. 2020. An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence*, 3, 4.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.