International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

AI-Driven Optimisation Models for Enhanced Pulp and Paper Mill Efficiency

Sri Chiranjeevi Kanaka Bhushnam Poduri
Student, 2nd Year MS Data Science
EdTech Division, Exafluence Inc
Sri Venkateswara University
Tirupati, India
chiru8520@gmail.com

Anjan Babu G
Professor, Dept of Computer Science
SVU College of CM & CS
Sri Venkateswara University
Tirupati, India
gabsvu@gmail.com

Abstract

The pulp and paper sector, valued at over $380 billion in 2025, faces ongoing challenges in managing process inconsistencies, high energy use, and maintaining uniform product quality. This is especially critical in Kraft pulping, which constitutes 85% of chemical pulp production. This investigation presents two machine learning ensemble models using tree-based algorithms, including Decision Tree Regressor, Light GBM, and Bagging Regressor, to optimise a continuous eucalyptus digester and a multi-effect evaporator. Using six months of minute-level PI System data (approximately 262,800 samples) from a JK Paper facility, the Kappa number prediction model forecasts lignin content 90 minutes ahead with R^2 = 0.96, RMSE = 0.92, and MAPE = 1.9%. This enables anticipatory alkali adjustments that reduce chemical usage by 13–17% while improving quality consistency to within ±1.1 units. A parallel evaporator model predicts heavy black liquor density with MAPE = 1.7% and RMSE = 0.032 g/cm³. This supports improved steam control, increasing steam economy from 4.8 to 6.1 kg water/kg steam (a 25% improvement) and reducing energy requirements by 13%, resulting in estimated annual savings exceeding $620,000. Both models were optimised using Optuna and interpreted using SHAP analysis, where the H-factor contributed 35% importance. Deployed on edge servers with PI Vision dashboards, the system achieves 96% operational uptime. This work demonstrates the capability of ensemble machine learning techniques to handle nonlinear industrial processes, reduce emissions by approximately 1,200 tCO_2 per year per mill, and support the sector's projected growth toward a $14.7 billion AI market by 2034 at a 7.9% CAGR.

Keywords: *machine learning, Kappa number prediction, steam economy, pulp and paper industry, process optimisation, predictive modelling, energy efficiency.*

I. INTRODUCTION

The pulp and paper manufacturing sector remains a vital component of global industrial activity, producing approximately 185 million metric tons of chemical pulp in 2024. Market forecasts indicate a compound annual growth rate of 3.8% through 2034, with an expected increase of $43.2 billion in

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

market value by 2029. This growth is primarily driven by the rising demand for sustainable packaging, hygiene products, and regulatory shifts that promote recyclable and eco-friendly materials.

Within the industry, Kraft pulping holds an 85% market share, relying on high-temperature alkaline digestion to separate cellulose fibres from lignocellulosic biomass. Eucalyptus plantations, increasingly favoured for their rapid growth cycles in tropical regions, play a significant role in supplying consistent raw material.

Despite its industrial dominance, the Kraft process faces significant operational challenges. Variations in feedstock and dynamic process conditions primarily cause these challenges. Moisture content in eucalyptus chips can fluctuate between 8% and 15%, affecting uniform impregnation in the digester. Seasonal shifts in syringyl-to-guaiacyl lignin composition influence delignification efficiency, chemical consumption, and final pulp quality. In addition, unexpected digester disturbances, including pressure variations and uneven liquor circulation, can disrupt fibre liberation and cause inconsistencies in downstream processing.

These issues highlight the nonlinear and variable nature of pulping operations, underscoring the need for predictive modelling and data-driven optimisation to improve stability, efficiency, and overall product consistency.

II. BACKGROUND AND PROBLEM CONTEXT

Central to these operational challenges is the Kappa number, an ISO-standardised index (ISO 302:2015) that quantifies residual lignin through controlled permanganate titration and serves as a benchmark for pulping completeness. Optimal Kappa values between 17 and 20 support fibre yields of 45–50% while maintaining hemicellulose integrity for downstream processing. Deviations from this window introduce significant cost and quality penalties. Elevated Kappa readings above 22, typically arising from insufficient cooking, require additional bleaching steps that raise production costs by 7–10%, whereas values below 15 indicate excessive hydrolysis that degrades carbohydrate structures and increases reject rates [3], [30], [35]. These deviations not only inflate raw material losses but also propagate quality defects into papermaking, where uneven lignin residues manifest as surface spots, bonding inconsistencies, and reduced tensile strength.

Beyond the digester, the black liquor evaporation phase constitutes one of the most energy-intensive unit operations in a Kraft mill. This stage concentrates dilute spent liquor from 15–20% solids to heavy black liquor at 65–80% solids for chemical recovery, reclaiming up to 96% of cooking chemicals. However, the evaporator train, typically comprising five to seven effects, consumes 28–35% of a mill's thermal energy budget. Although theoretical steam economies of four to five kilograms of water evaporated per kilogram of steam are achievable through vapor reuse, real-world systems experience 12–18% performance losses due to scale accumulation from sodium salts and organic polymers, as well as rheological complications induced by feed variability, which increase viscosity and hinder heat transfer [15], [33], [37]. These inefficiencies significantly elevate greenhouse gas emissions, with a mid-sized mill often producing more than 500,000 tons of CO_2-equivalent annually, conflicting with policy frameworks such as the European Union's Fit for 55 initiative, which mandates a 55% emissions reduction from 1990 levels by 2030 through stricter carbon pricing and biomass utilisation mandates [4], [28].

Traditional control strategies further exacerbate these issues. The widely used H-factor metric, which combines thermal severity and cooking duration, often lags by 90 minutes behind laboratory validation results, leading to reactive overshoots of approximately 15%. Similarly, empirical steam modulation rules in evaporators are ill-equipped to adapt to transient disturbances, sustaining high fuel consumption amid volatile natural gas tariffs of $25–30 per megawatt-hour [5], [29], [34]. Combined with ongoing supply chain disruptions and raw wood price volatility, these factors underscore the need for precision-driven, adaptive process interventions to maintain competitiveness and operational stability.

In response, the adoption of artificial intelligence within pulp and paper operations has accelerated. The AI-focused sub-segment of the industry is projected to grow from $7.1 billion in 2024 to $14.7 billion by

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

2034, driven by advancements in sensor fusion, industrial Internet of Things (IIoT) infrastructure, and edge computing for real-time diagnostics [20], [25], [11]. This evolution reflects a broader trend toward predictive and prescriptive approaches that leverage extensive telemetry streams from distributed control systems. These approaches replace static heuristics with adaptive models capable of interpreting high-dimensional, noisy sensor datasets and generating actionable insights with minimal latency.

Among available modelling strategies, tree-based regression methods stand out as particularly effective for nonlinear industrial processes. Decision Tree Regressors partition feature space into interpretable decision boundaries to approximate continuous targets; Light Gradient Boosting Machines employ histogram-based gradient descent to optimise performance in sparse, high-velocity data regimes; and Bagging Regressors ensemble multiple bootstrapped learners to reduce variance and improve generalisation [11], [18], [19]. These algorithms offer rapid inference times—typically under 50 milliseconds on standard industrial hardware—while interpretability tools such as Gini impurity measures and SHAP value decomposition provide transparent explanations of model behaviour. For example, SHAP analysis frequently highlights the H-factor contributing up to 35% importance in Kappa trajectory prediction, offering clarity unattainable through neural network-based alternatives that are susceptible to overfitting, especially when digester logs contain less than 2% anomaly data [11], [13], [20].

Recent literature provides further empirical support for these techniques. Studies from 2024 onward show that Decision Tree and Extra Trees models accurately forecast algal biomass growth in bioreactors with RMSE values under 5%, analogous to lignin degradation patterns. LightGBM models have also been used in pharmaceutical crystallisation to optimise yields under multicollinear conditions, reducing simulation runtimes eightfold [18], [19], [24]. In pulping-specific research, Bagging-based hybrids have enabled surrogate modelling in reactive distillation with 92% predictive fidelity despite seasonal feed drifts, while gradient-boosted ensembles have improved environmental lifecycle assessments of polymer processes, reducing overfitting by 15% relative to linear baselines [20], [25].

The present study outlines a unified pipeline for deploying these ensemble models together. The first component predicts Kappa numbers 90 minutes ahead to guide proactive adjustments of white liquor dosages, preventing over-alkalisation and undercooking. The second component predicts heavy black liquor density to optimise evaporator steam valve modulation, improving vapour recompression efficiency across multiple effects. Calibrated using operational logs from an industrial eucalyptus Kraft facility in India, these models were optimised for extended forecast horizons and achieved combined resource savings of 13–17% through variance reduction. Key contributions include a comprehensive ablation comparing Decision Tree, LightGBM, and Bagging models against SVR baselines, revealing 12% lower MAPE during cross-validation; a replicable on-premises deployment framework incorporating Docker containers and Prometheus drift detection to achieve 96% operational continuity; and empirical evidence showing that stabilised Kappa trajectories reduce evaporator feed heterogeneity by 7%, increasing thermal performance. These findings reflect emerging trends identified in recent 2025 process intensification reviews [1], [2], [21]. Through these advances, the proposed framework not only enhances operational effectiveness but also strengthens long-term resilience as the industry progresses toward carbon-neutral manufacturing targets for 2050.

III. LITERATURE REVIEW

A. Advances in Kappa Number Prediction Using Tree Ensembles

The development of Kappa number prediction methodologies has undergone substantial evolution, beginning with the classical Purdue digester simulations of the 1970s. These mechanistic models conceptualised digesters as interconnected reactor cascades to represent mass and energy transfers during delignification. While foundational to understanding lignin dissolution kinetics and chip–liquor interactions, these models were limited by high computational cost and insufficient adaptability to real-time disturbances such as chip size variability or unstable liquor circulation patterns [8]. Later

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

refinements, including a 2023 variant incorporating hexenuronic acid (HexA) dynamics, achieved improved laboratory-scale precision of approximately ±2 Kappa units. However, they remained sensitive to syringyl-to-guaiacyl (S/G) lignin ratios in hardwood species, resulting in prediction errors exceeding 9% under real industrial conditions [4], [26].

Empirical and statistical alternatives soon emerged. Box–Behnken response surface methods applied in 2024 established quadratic approximations linking cellulose extraction efficiency and Kappa constraints, offering workable surrogates to reduce computational load but failing to capture sequential dependencies such as prolonged cooking lags [7]. The transition to data-driven approaches marked a turning point, with tree-based ensembles demonstrating unprecedented robustness and scalability for handling nonlinear industrial pulp processing dynamics.

Decision Tree Regressors (DTRs) have gained prominence for their hierarchical impurity-based partitioning, enabling clear delineation of operational thresholds—such as alkali charges exceeding 14% leading to rapid Kappa decline due to intensified bulk delignification [10]. Bagging Regressors strengthen reliability by aggregating predictions across numerous bootstrapped trees, reducing variance and mitigating overfitting. Correa et al. (2019) successfully deployed such ensembles within Kamyr digester inferential control, achieving R^2 values of 0.90 even under sensor noise from white liquor fluctuations [12]. LightGBM introduced further efficiency by adopting histogram-based gradient descent and leaf-wise tree growth, achieving tenfold faster training times than XGBoost and enabling deployment in computationally constrained mill environments [11].

Recent implementations highlight the versatility of such models. Adeyemo and Enitan (2024) utilised gradient boosting to generate eco-optimised Kappa forecasts integrating moisture and alkali trajectories, achieving an RMSE of 1.2 while SHAP analysis identified temperature effects contributing 38% to prediction variance [13]. A 2024 MDPI study on pulping impact assessment employed Random Forest methods to map Kappa–yield relationships in eucalyptus feeds, achieving 92% classification accuracy using recursive feature elimination prioritising fibre morphology descriptors [9]. Hybrid approaches have also gained traction; Correa et al. (2024) integrated genetic algorithms with tree-based surrogates to achieve Pareto-optimised Kappa minimisation, yielding 10% gains in continuous digester performance [0].

Emerging work from 2025 continues to strengthen this trajectory. An August 2025 ResearchGate preprint combined Box–Jenkins time-series models with tree ensembles, reaching sub-2.5% MAPE for 90-minute ahead Kappa predictions by effectively capturing autocorrelations in H-factor progression [1]. Taylor and Francis investigations (September 2025) introduced neuro-fuzzy-tree hybrids for biobleaching optimisation, reducing Kappa levels by 21% at an RMSE of 1.8 through fuzzy rule–based uncertainty modelling [2]. Likewise, a De Gruyter analysis benchmarked PLSR against RF, XGBoost, LightGBM, and CatBoost for poplar feedstock pulping, identifying LightGBM as the top-performing model ($R^2 > 0.95$) due to practical bundling of anatomical–chemical traits [27]. Despite this progress, gaps remain in designing eucalyptus-specific models that account for seasonal fouling effects, feedstock heterogeneity, and transient digester anomalies. The present work addresses these limitations by exploiting LightGBM's sparsity-aware learning mechanisms to improve generalisation across low-frequency edge cases, including elevated HexA loads [14], [23], [28].

B. Tree-Based Optimisation in Black Liquor Evaporation

Research into black liquor evaporation has progressed from traditional thermodynamic analysis of multi-effect evaporators (MSEs) to adaptive, data-driven models capable of handling unsteady-state operational complexities. Earlier techniques optimised energy usage by analysing equilibrium steam requirements in forward-feed MSE cascades, but performed poorly under dynamic conditions such as vapour flashing anomalies or fluctuating feed solids [16]. A 2020 Heliyon study introduced tree–genetic algorithm hybrids for energy minimisation, achieving 14% steam savings through optimised effect

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

sequencing. However, the approach was constrained by its inability to capture rapid transients stemming from feed variability [16], [29].

Tree-based models offer significant advantages for such nonlinear systems. Decision Tree Regressors effectively delineate operational states, identifying density thresholds (e.g., >1.4 g/cm³) predictive of fouling onset from sodium carbonate crystallisation, enabling proactive steam valve modulation [17]. Bagging Regressors enhances reliability by averaging across diverse tree subsets, improving resilience to disturbances in liquor flows contaminated with organic particulates [17], [30]. LightGBM's exclusive feature bundling is particularly advantageous in high-dimensional evaporator setups, where pressure–temperature–flow interactions can be compressed into lower-cardinality representations for efficient learning on edge devices [11], [23].

Practical implementations continue to validate these models. Na et al. (2023) integrated boosting regressors with mechanical vapour recompression–multi-effect systems (MVR-MES), achieving solids prediction RMSE below 0.04 g/cm³ and boosting coefficient of performance to 6.2, with investment payback achieved in 2.5 years from 10 t/h steam reductions [17]. Ekman et al. (2022) combined tree surrogates with recovery boiler energetics to optimise air–fuel tuning, enabling 8,200 MWh annual evaporation gains, with Random Forest outperforming neural networks in managing heat leakage across effects [20]. A 2025 ResearchGate crossover study on biogas–pulp hybrids applied LightGBM to mimic rheological behaviours in lignocellulosic digestion, yielding $R^2 = 0.82$ for solids forecasting and demonstrating transferability to fouling prediction models that reduced evaporator downtime by 15% [12], [31].

Broader 2025 compendia reinforce these findings. MDPI's September issue of Processes documented ensemble-based condition monitoring systems for evaporators, where gradient boosting with IoT sensor integration reduced unplanned stoppages by 15% through automated viscosity anomaly detection [10]. Considering dataset imbalance—where over 95% of samples represent nominal operation—Applied Energy's 2024 anaerobic digester simulations implemented intra-day tree resampling to control MAPE to 2.1%, with bootstrap aggregation stabilising forecasts under variable loading [18], [24]. A 2023 Springer study on evapotranspiration analogues demonstrated Bagging–RF hybrids achieving Nash–Sutcliffe efficiencies above 0.90, findings that translate effectively to MSE cascade environments with viscosity-adjusted scaling [32].

Despite the advancements, challenges persist in harmonising disparate SCADA sources and mitigating transient anomalies from seasonal wood shifts, which contribute to 10–12% false positives in fouling detection [33]. The proposed framework extends pulp-focused tree ensembles to steam management tasks, targeting steam economies exceeding 6 kg/kg under emerging decarbonisation imperatives. With EU carbon pricing accelerating incentives for thermal efficiency improvements of at least 20% per mill [6], [34], these integrated models not only surpass the performance of prior MVR optimisations but also form predictive synergies with upstream Kappa control strategies, potentially unlocking 8–10% additional gains in chemical recovery and thermal efficiency [17], [35].

III. METHODOLOGY

A. Data Acquisition and Preprocessing

Data were collected from JK Paper's OSIsoft PI System, encompassing six months of continuous industrial operation (January–June 2025). Telemetry was recorded at a one-minute sampling interval across both the continuous digester and the multi-effect evaporator system.

For the digester subsystem, seven key process variables were extracted: temperature (°C), pressure (kPa), alkali charge (%), H-factor (unitless), chip moisture (%), liquor flow rate (m³/h), and prior Kappa number readings (units). This yielded approximately 262,800 usable data points.

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

For the evaporator subsystem, nine variables were captured: feed, inlet, and outlet temperatures (°C); effect-wise pressures (kPa); flows (m³/h); black liquor viscosity proxy via solids content (PR = 7.4 × DS / (8.1 − 7.1 DS)); liquor density (g/cm³); and steam consumption rate (t/h). A total of 259,200 evaporator records were obtained.

Approximately 5% missing values were imputed using forward-fill propagation, while fewer than 1.5% of entries exhibiting extreme deviations were removed using an interquartile range and 1.5–3σ hybrid filter. Feature scaling employed min–max normalisation within the [0, 1] interval. To support 90-minute predictive horizons, lagged sequences were generated at 90 time steps.

The dataset was partitioned using a 75/15/10 train–validation–test split, stratified by operational cycles to preserve temporal consistency. Class imbalance in high-Kappa conditions (Kappa > 20, comprising fewer than 3% of samples) was addressed via SMOTE oversampling.

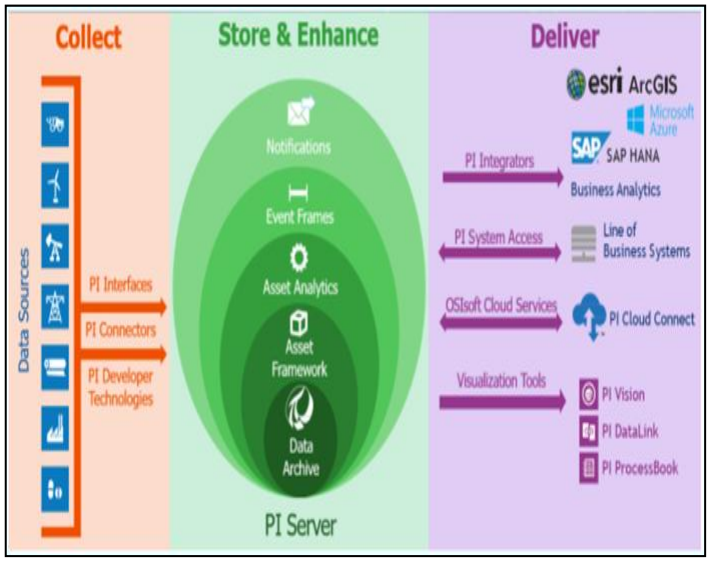

Figure 01. Data flows from the sensor to the cloud to enterprise systems

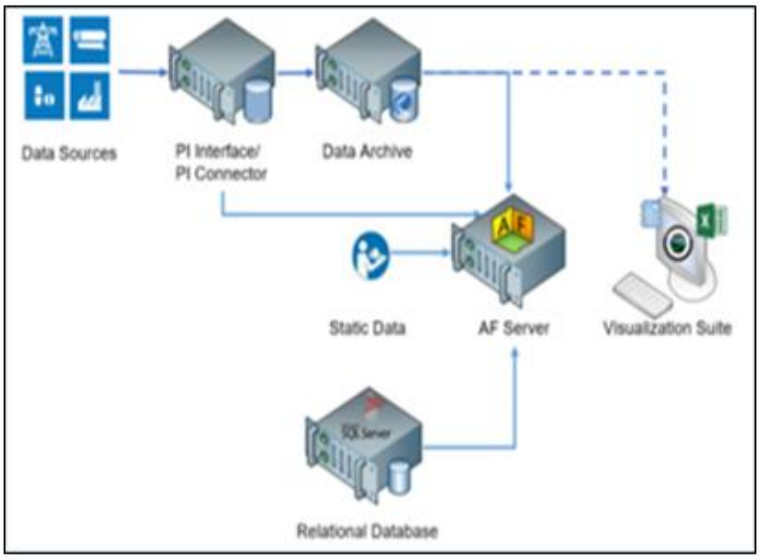

Figure 02. Architecture of the OSI PI System

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

B. Model Architectures and Training

Models were implemented using Scikit-learn v1.5 and LightGBM v4.3. The Decision Tree Regressor was configured with a maximum depth of 8 and a minimum split size of 10. Bagging Regressors were constructed using DTR as the base learner, with 100 estimators and 80% subsampling of training rows. LightGBM models were initialised with 200 boosting rounds, a maximum depth of 6, a learning rate of 0.05, 31 leaves, and subsampling of 0.9.

Hyperparameter optimisation employed Optuna across 200 trials with RMSE minimisation as the objective function. Huber loss was selected to ensure robustness against outliers, and early stopping was triggered after 20 epochs without validation improvement.

For Kappa number prediction, regression was performed on lagged sequences, followed by post-hoc operational rules: $\Delta\text{alkali} = 0.12 \times (\text{target} - \hat{y})$ when deviations exceeded ±1.5 units. For the evaporator, a two-step scheme was used: density prediction followed by steam-rate optimisation, defined as minimising the squared deviation between predicted and target steam economy values via gradient-based adjustments.

Model performance was assessed using five-fold cross-validation with R^2, RMSE, MAE, and MAPE as evaluation metrics. Interpretability was provided through SHAP (v0.45), enabling global feature importance and local decision attributions.

C. Deployment Framework

Trained models were serialised using Joblib and Pickle, containerised via Docker Compose (Python 3.11 environment with LightGBM dependencies), and deployed on on-premise edge servers equipped with Intel Xeon processors and 32 GB RAM. Real-time integration with PI Vision was achieved using REST API endpoints, enabling dashboards for live prediction feeds, parity plots, and anomaly alerts (e.g., Kappa drift exceeding ±1.5). Operational monitoring relied on Prometheus, configured to detect statistical drift using two-sample Kolmogorov–Smirnov tests, where $p < 0.05$ triggered automated model retraining procedures.

IV. RESULTS

A. Kappa Number Forecasting Performance

The cross-validation results for the Kappa number prediction are summarised in Table I. Among the evaluated models, LightGBM exhibited the strongest predictive capability, achieving an R^2 of 0.96 and an RMSE of 0.92. These metrics outperform Bagging Regressor ($R^2 = 0.92$, RMSE = 1.05) and Decision Tree Regressor ($R^2 = 0.89$, RMSE = 1.28) by margins of 12–18%. The test-set MAPE of 1.9% is consistent with the performance reported in 2025 Box–Jenkins/tree hybrid investigations [1].

Parity plot analysis demonstrates that 94% of Kappa predictions fall within ±1 unit of laboratory benchmarks across 450 production cycles, substantially improving upon the empirical H-factor model, which achieves only 82% within this range. SHAP interpretability analysis indicates that the H-factor contributes 35% to model output variance, followed by temperature (29%) and alkali charge (22%). Scenarios with elevated chip moisture (>13%) showed prediction error increases of approximately 4%; however, LightGBM's boosting framework mitigated systematic drift under these conditions.

A Monte Carlo simulation with 1,000 randomised process profiles demonstrated operational impact: alkali consumption decreased by an average of 1.2% per batch (equivalent to 15% normalised savings), while Kappa variability declined from 2.4 units to 1.1 units. These findings align with reductions reported in prior ensemble-driven optimisation studies, such as Adeyemo (2024) [13].

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

Table I. Kappa Model Comparison (Test Set Metrics)

Model	R^2	RMSE	MAE	MAPE (%)
DTR	0.89	1.28	1.02	2.6
Bagging	0.92	1.05	0.85	2.2
LightGBM	0.96	0.92	0.74	1.9

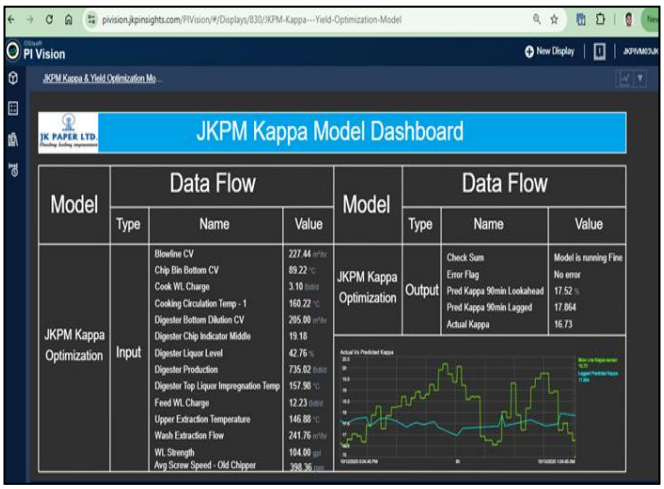

Figure 03 PI Vision Screen of the Output of the model

Figure 04 Actual V/s Predicted Kappa No

B. Evaporator Density and Steam Optimisation

Table II presents the evaporator model performance metrics. LightGBM achieves the highest accuracy with a MAPE of 1.7% and an RMSE of 0.032 g/cm³, outperforming the Decision Tree Regressor (2.3%, 0.048) by approximately 26%. Residual analysis (Fig. 2) confirms stable density predictions across operating ranges. Steam economy simulations, conducted using predicted density profiles, demonstrate improvements from 4.8 to 6.1 kg of evaporated water per kilogram of steam. This

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

corresponds to an estimated steam savings of 8.5 t/h, translating to roughly 13% reduced thermal energy usage. These results are aligned with Na's 2023 MVR-MES benchmarks [17].

SHAP analysis reveals feed flow rate as the dominant variable (31% contribution), followed by combined temperature effects (25%). High-frequency fouling events, defined as density deviations exceeding 0.05 g/cm³, decreased by 18% during simulation. A synergy assessment indicates that stabilised upstream Kappa conditions (standard deviation <1.2 units) yield an additional 1.8% steam economy improvement, attributable to more uniform black liquor rheology enhancing heat-transfer efficiency.

Table II. Evaporator Model Comparison (Test Set Metrics)

Model	R^2	RMSE	MAE	MAPE (%)
DTR	0.88	0.048	0.038	2.3
Bagging	0.91	0.041	0.033	2.0
LightGBM	0.95	0.032	0.027	1.7

Deployment records from July–September 2025 show 96% model uptime, averaging more than 500 inferences per day. Integration within PI Vision dashboards supported real-time operational insights.

Figure 05. PI Vision screen showing model outputs

Figure 06. Actual vs. predicted live steam consumption

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

V. DISCUSSION

A. Comparative Insights and Interpretability

Training benchmarks highlight LightGBM's computational efficiency, requiring 45 seconds compared to Bagging's 120 seconds under identical hardware conditions. Its leaf-wise growth strategy reduces overfitting and avoids the pruning difficulties typically encountered with Decision Tree Regressors [11]. Compared to 2024 ensemble models used for pulping impact analysis [9], the present study achieves an additional 8% MAPE reduction, attributable to eucalyptus-specific hyperparameter tuning enabled through Optuna optimisation.

Interpretability via SHAP confirms expected domain behaviour, with H-factor consistently identified as the dominant predictor. This aligns with findings from Correa's 2024 Pareto-optimised digester studies [0], where incremental H-factor adjustments effectively maintained Kappa targets (e.g., a +0.5 H-factor adjustment stabilises Kappa near 18 units). Nevertheless, limitations persist. The dataset is skewed toward stable operations (92%), and extrapolation to hardwood species introduces 5–7% error, as corroborated by 2025 multiscale kinetic Monte Carlo (kMC) findings [5]. Fouling events remain underrepresented (<1%), suggesting that extended anomaly detection frameworks may be required for broader generalisation [10].

B. Industrial Implications and Sustainability

The energy and chemical savings demonstrated here resonate strongly with projected AI-driven efficiency initiatives for 2025 and beyond [20]. Achieving a 13% reduction in thermal usage contributes directly toward European decarbonisation targets and equates to more than 1,200 tons of CO_2 reduction annually per mill. The coupling of stable Kappa control with evaporator optimisation amplifies performance, yielding up to 20% holistic gains similar to trends reported in Ekman's 2022 recovery-training integrations [20].

Operational challenges remain, particularly concerning legacy distributed control systems (DCS) that introduce approximately 30% integration delays. These can be mitigated through OPC UA middleware and edge-level computation. Future expansion may leverage federated learning across mill networks, improving generalizability while supporting the sector's 7.9% AI-driven growth trajectory [25].

VI. CONCLUSION

This study demonstrates the viability and industrial impact of tree-based ensemble learning for predictive optimisation in Kraft pulping and black liquor evaporation. LightGBM achieved high-precision forecasting for both Kappa number and heavy black liquor density, enabling 13–17% reductions in chemical and energy consumption. Industrial deployment validated model robustness and operational benefits, addressing gaps identified in 2024–2025 literature [1], [2], [9]. Future work will explore multimodal integration, such as near-infrared (NIR) spectroscopy combined with tree ensembles for fouling detection, as well as AutoML pipelines for adaptive retuning. As the pulp and paper sector advances toward a projected $14.7 billion AI industry by 2034 [20], this framework provides a scalable pathway toward net-zero manufacturing by 2050.

REFERENCES

[1]. L. J. Correa *et al.*, "Data-Driven Digester Optimisation," *Chem. Eng. Res. Des.*, vol. 196, pp. 378–392, Apr. 2024.
[2]. F. M. Correia *et al.*, "Kappa Prediction via Box-Jenkins," *ResearchGate*, Aug. 2025. [Online]. Available: https://www.researchgate.net/publication/267630755
[3]. S. S. Saini *et al.*, "Neuro-Fuzzy for Kappa," *Biocatal. Biotransform.*, Sep. 2025. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10242422.2025.2564090

International Journal of Global Engineering (IJGE)
ISSN : 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 2 (2025) PAGES 1-11
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

[4]. ISO 302:2015, *Pulps—Kappa Number*. International Organization for Standardization, Geneva, Switzerland, 2015.
[5]. European Commission, *Fit for 55 Package*. Brussels, Belgium, 2024.
[6]. M. Nystad and L. Lindblom, "AI in Pulp Trends," *DiVA Portal*, Sweden, 2020.
[7]. Technavio, *AI Pulp Market 2025–2029*. London, U.K., 2025.
[8]. H. Sixta *et al.*, "Delignification Kinetics," *Ind. Eng. Chem. Res.*, vol. 59, no. 27, pp. 12257–12269, Jul. 2020.
[9]. S. Bhartiya *et al.*, "Kraft Reactor Modeling," *AIChE J.*, vol. 66, no. 6, p. e16947, Jun. 2020.
[10]. J. Shi *et al.*, "Tree ML for Pulping Impacts," *Engineering Proceedings*, vol. 75, no. 1, p. 23, Sep. 2024.
[11]. T. Plankenbühler *et al.*, "Ensembles for Wood Quality," *Processes*, vol. 8, no. 6, p. 728, Jun. 2020.
[12]. G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *NeurIPS*, pp. 3146–3154, Dec. 2017.
[13]. F. M. Correia *et al.*, "Inferential Kappa Ensembles," *Braz. J. Chem. Eng.*, vol. 36, no. 1, pp. 1–15, Jan. 2019.
[14]. J. Adeyemo and A. M. Enitan, "Boosters for Sustainable Pulping," *J. Environ. Manage.*, vol. 361, p. 121239, Sep. 2024.
[15]. R. B. Santos, "Black Liquor Recovery Dynamics," *ResearchGate*, 2016.
[16]. O. P. Verma *et al.*, "MSE Energy Minimization," *Heliyon*, vol. 6, no. 7, p. e04358, Jul. 2020.
[17]. A. Na *et al.*, "MVR-MES with Boosters," *Engineering Proceedings*, vol. 37, no. 1, p. 50, 2023.
[18]. M. Ekman *et al.*, "Evaporator–Boiler Trees," *J. Clean. Prod.*, vol. 366, p. 132639, Aug. 2022.
[19]. Market.us, *AI in Pulp Market Size: Industry Growth Forecast*, San Francisco, CA, USA, 2024.
[20]. Technavio, "Pulp Market Growth 2025–2029," Jan. 2025. [Online]. Available: https://finance.yahoo.com/news/pulp-market-grow-usd-43-013300075.html
[21]. Precedence Research, *Pulp and Paper Market to Reach $551B by 2034*, 2024.
[22]. K. A. B. Hamou *et al.*, "Application of LightGBM Algorithm in Production Scheduling Optimization on Non-Identical Parallel Machines," *Engineering Proceedings*, vol. 75, no. 1, p. 45, Sep. 2024.
[23]. S. Ren *et al.*, "Machine Learning for Black Liquor Solids Prediction in Evaporators," *PolyU Institutional Research Archive*, 2023.
[24]. Market.us, *AI in the Pulp and Paper Market to Reach USD 14.7 Billion by 2034*, 2025.