ISSN: 0379-5462
IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association
Vol. 38, E2, 2026, p. 115-120

DATA BEFORE AI: A Practical Case Study in Paper Mill Data Foundation Building

Twinkle Varu*
Technical product marketing Analyst

Aniket Gaikwad*
Associate Product Manager - Digital Solution

Jayesh Bachhav*
Senior Product Manager

**HABER*

Abstract: *The pulp and paper industry is increasingly exploring artificial intelligence (AI) and advanced analytics to improve quality consistency, resource efficiency, sustainability, and operational performance. However, the effectiveness of AI depends less on advanced algorithms and more on the reliability and contextual quality of plant data. This paper presents a practical case study demonstrating how structured data foundations can enable AI-ready manufacturing in pulp mills. A data foundation programme was implemented at a leading pulp mill by standardizing manual measurements, laboratory testing, operator checks, and process data into a unified digital environment with proper timestamps and process context. Using this structured dataset, an 8-month Golden Batch Analysis was conducted on 4000 digester batches across six vessels. The study revealed that only 23% of batches simultaneously achieved the target Kappa range (14–16) and Viscosity range (220–260) s, highlighting significant process variability. The analysis identified the operating conditions associated with consistently successful batches and demonstrated how structured plant data can uncover actionable process insights. The study shows that AI readiness in pulp mills can begin with existing operational practices when data is systematically captured, contextualized, and analyzed. For the Indian paper industry, this approach offers a practical pathway toward future-ready manufacturing and advanced process optimization*

Keywords: *Data foundation, golden batch, pulp quality, digester control, Kappa, Viscosity, AI readiness, sulphite pulping, process optimisation*

1. Introduction

The Indian pulp and paper industry stands at a critical juncture. Rising wood and fibre costs, tightening environmental regulations, increasing competition from imports, and customer demands for consistent quality are converging simultaneously. At the same time, AI and advanced process analytics are no longer confined to petrochemicals or semiconductors - they are increasingly being deployed, or at least discussed, in paper mills around the world [1].

Yet a persistent gap exists between the promise of AI-driven manufacturing and its realisation on the mill floor.[1] Most discussions focus on algorithms: predictive models, neural networks, reinforcement learning. What receives far less attention is the precondition that determines whether any of these tools can function at all - the quality, continuity, and contextual richness of the underlying process data.

A predictive model for Kappa number, however sophisticated, is only as good as the Kappa measurements it trains on [2, 3]. If those measurements are recorded inconsistently, lack timestamps, or cannot be linked to the grade being produced, the shift operating, or the chemical dosages applied, the model has nothing to learn from. The algorithm is not the bottleneck. The data is.

This paper argues that data foundation building - the systematic capture, validation, and contextualisation of process data - is the necessary and often-skipped first step for every mill pursuing AI [2, 3]. It then demonstrates what becomes possible once that foundation exists, through a detailed case study of the Golden Batch Analysis at a pulp mill: an 8-month retrospective study of 4000 digester batches across 6 vessels, which revealed a 23% joint Kappa -Viscosity success rate and a clear, actionable operating envelope for improving it.

2.The Data Challenge in Indian Paper Mills

2.1 The Gap Between Data Generation and Data Utility

Pulp and paper mills generate substantial process data every day: manual measurements of chip moisture, cooking chemical concentrations, Kappa, Viscosity, and brightness; continuous signals from

equipment; shift logs; and laboratory quality outcomes.[2, 4] Despite this apparent richness, most data is either lost or rendered analytically unusable before it can be leveraged. The failure modes are consistent across mills:

- Manual measurements recorded on paper or unstructured spreadsheets, with no standardised format and no timestamp beyond the date
- Laboratory results stored in standalone systems that cannot be linked to the batch or process conditions that produced the sample
- No contextual metadata: measurements not tagged with grade, furnish composition, operating shift, or chemical dosages applied
- Inconsistent sampling cadence: some variables measured hourly, others daily, others only when a problem is suspected
- No unified data model: process data, laboratory data, and manual records in separate silos that cannot be joined for analysis

The result is a situation where a mill may hold years of historical records and yet be unable to answer basic questions: What were the operating conditions during the 50 batches that achieved the best Kappa - Viscosity combination last quarter? How does chip moisture affect viscosity at different cooking temperatures? These are not difficult analytical questions. They are impossible questions - not because the answers do not exist in the data, but because the data cannot be assembled into a form that supports the analysis [1, 4].

2.2 Primacy of Data Quality Over Algorithmic Sophistication

The temptation when deploying AI in a mill is to begin with the model. [1] Mill managers and technology vendors focus on the sophistication of the analytics: neural networks, time-series forecasting, multivariate process control. This is understandable - the model is the visible, exciting part of the system.

But a model trained on incomplete, poorly timestamped, or context-free data will learn the wrong things [1]. It will identify spurious correlations, miss true causal relationships, and produce recommendations that operators rightly distrust. Once operator trust is lost, even a technically capable system goes unused.

The sequence matters. Data discipline must precede model sophistication. A simple statistical model trained on clean, well-contextualised data will consistently outperform a sophisticated model trained on poor data. The Indian paper industry, in its understandable enthusiasm for AI, frequently inverts this sequence - deploying technology before the data substrate is ready to support it.

2.3 The Indian Context

Several features of the Indian paper industry make the data challenge particularly acute. Fibre variability is higher than in Scandinavian or North American mills operating on homogeneous plantation timber: Indian mills handle agricultural residues, recycled fibre, and mixed hardwood–softwood furnishes,[2, 3] each with different moisture content and cooking chemical demand. This variability amplifies the importance of contextual data capture, because the same cooking recipe produces different results on different furnishes.Process complexity is also a factor: many Indian mills operate older equipment with limited online instrumentation, making manual data capture especially critical.[5] At the same time, sustained margin pressure means that the optimisation opportunities enabled by good data - reduced chemical consumption, higher throughput, lower reject rates - represent real and measurable value. The cost of poor data is not abstract; it appears in every batch that misses specification.

3. Building a Data Foundation: A Structured Approach

3.1 Starting Without Large-Scale Sensor Deployment

A common misconception is that data foundation building requires significant capital investment in online instrumentation - DCS upgrades, new analyser installations, sensor network deployments. While these investments may eventually be valuable, they are not the necessary starting point. Most mills already have access to more useful data than they are capturing effectively. The initial opportunity lies in disciplining and digitising the manual measurement and laboratory routines that already exist.

A data foundation programme at a leading pulp mill demonstrated this clearly. The mill had limited online instrumentation, but generated rich manual data through routine sampling, laboratory testing, operator checks, and quality observations.

3.2 The Five-Step Framework

The steps enlisted below are not a technology implementation checklist. They are a data discipline programme - applicable to any mill regardless of instrumentation level - that transforms latent process data into an analytically usable form. Each step builds on the previous; skipping or compressing any one of them degrades the quality of what follows.

- Step 1 – Define Critical Variables: Not every measurement matters equally. With process engineering input, identify the variables that most influence key quality outcomes - in a pulp mill, typically Kappa, Viscosity, brightness, and reject rate. This scoping focuses collection efforts and avoids the common failure of collecting everything and analysing nothing effectively.
- Step 2 – Standardise Measurement Frequency: For each critical variable, define how often it should be measured and by whom. Inconsistent intervals - a variable measured hourly by one shift and twice daily by another - make time-series analysis impossible.
- Step 3 – Digital Capture at Source: Every measurement was captured digitally, at the point of measurement, with an accurate timestamp. Paper logs and after-the-fact spreadsheet entry were not acceptable substitutes. Mobile data capture tools, even simple ones, digitised manual measurements at the point of collection. Critically, data capture must also occur consistently at the same physical location each time. Readings taken at varying positions along a process line - even for the same variable - introduce systematic noise that undermines comparability. An additional benefit of fixing the capture location is forward compatibility: where possible, the manual sampling point should be co-located with the intended installation point of any future online sensor or analyser. This alignment ensures that the historical manual dataset and the future instrument signal are directly comparable, enabling seamless transition from manual to automated measurement without loss of data continuity.
- Step 4 – Contextualise Every Measurement: Each measurement was tagged with the process context in which it was taken: grade being produced, furnish composition, operating shift, chemical dosages applied. Without this context, a Kappa reading of 9.2 is merely a number; with context, it becomes a data point comparable to all other readings taken under similar conditions.
- Step 5 – Integrate into a Unified Data Layer: Manual, laboratory, and historian data were integrated into a unified operational intelligence layer. This layer standardized process context, batch lineage, quality outcomes, and operating conditions into a common schema that could support analytics, process optimization, and future AI applications - the data layer on which all subsequent analytics would operate. Its design anticipated the use cases it needed to support: process variability analysis, root cause identification, chemical optimisation, quality prediction, and eventually closed-loop control. The digital layer also encompasses a process knowledge graph linking batches (Figure 1), quality outcomes, operating conditions, and equipment context.

3.3 Operator Adoption and Change Management

The technical steps above are necessary but not sufficient. The most important determinant of success is operator adoption - the degree to which the people who generate the data actually use the new capture routines consistently and accurately. Operators must understand why the data is being collected and

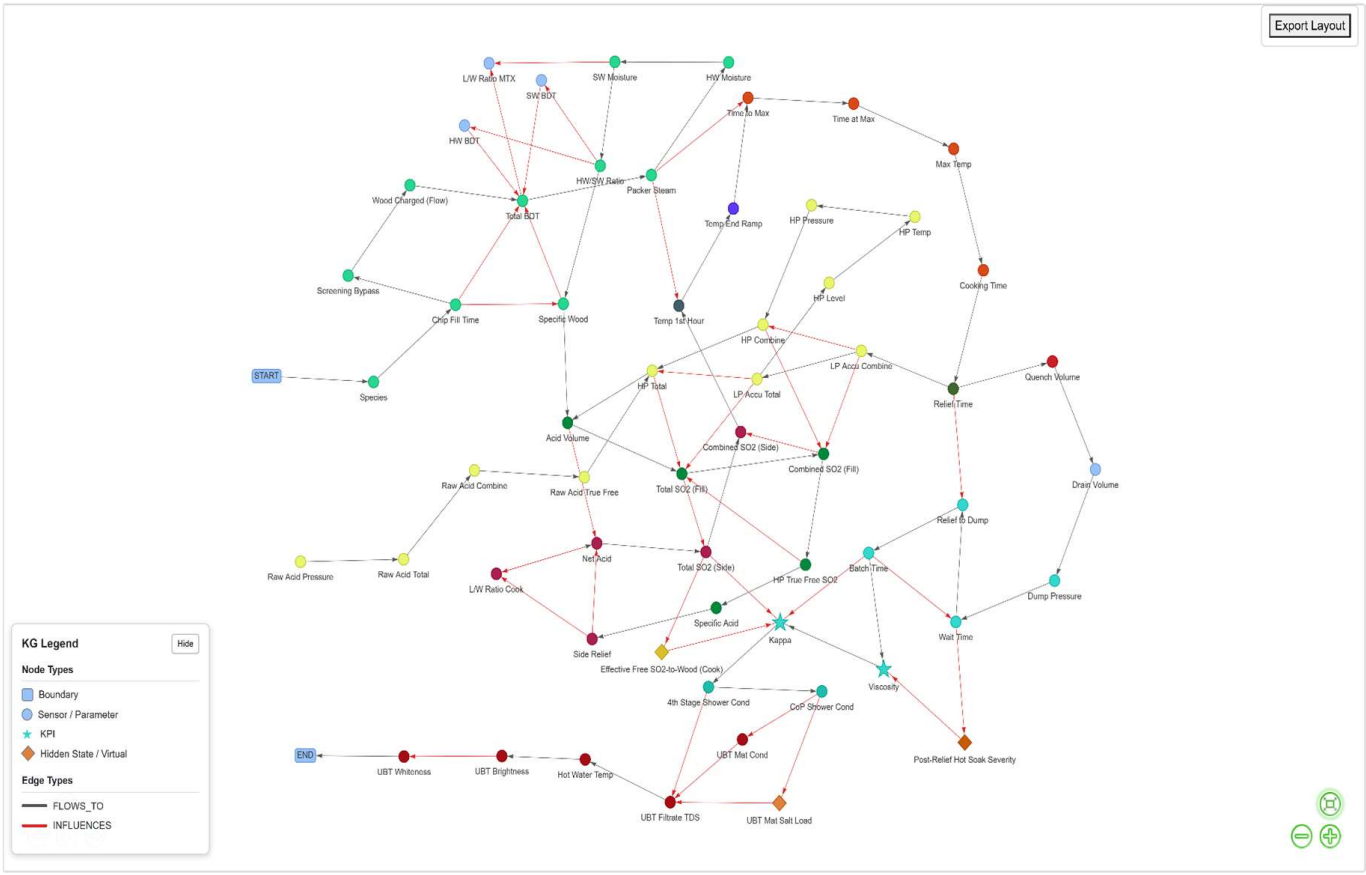

Figure 1: Process knowledge graph

how it will be used; tools must make capture easier, not harder, than the previous method; and operators must see their data producing insights that are genuinely useful to them on shift.

When operators see their data leading to recommendations that improve their shift performance, the virtuous cycle of data quality and analytical insight reinforces itself. The three-week programme at the pulp mill achieved this through co-design of measurement routines with operators, daily feedback sessions where early data patterns were shared, and a deliberate focus on immediately practical use cases such as batch quality prediction. (Figure 2)

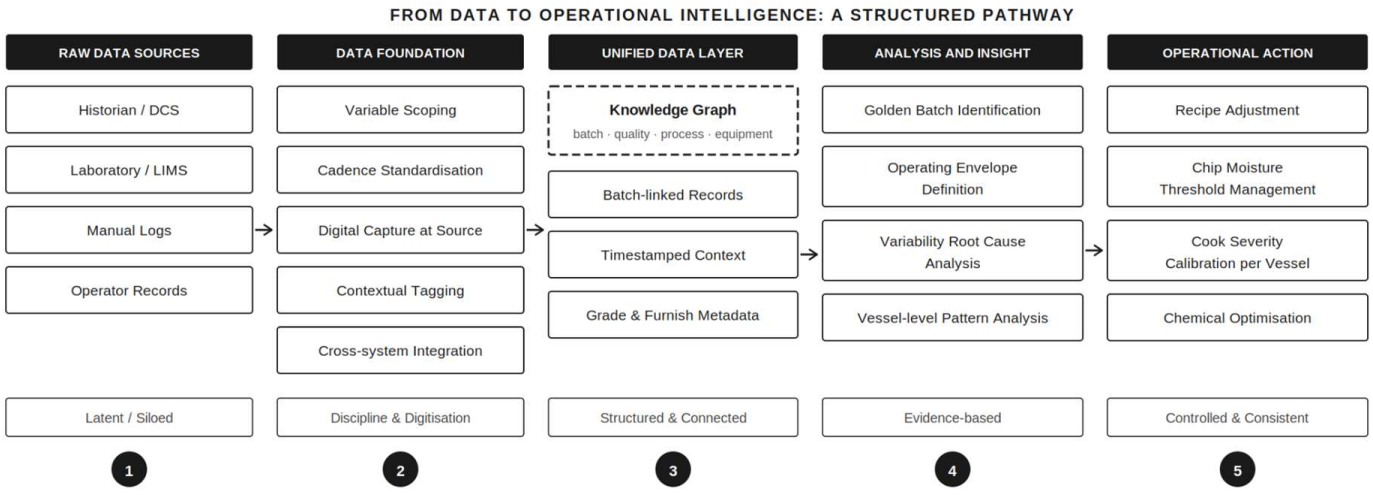

Figure 2: Framework for AI-Ready Data and Operational Intelligence in Paper Mills

The Mt. Fuji Logbook (Figure 3) captured shift-wise manual entries for critical process parameters - Kappa Number, Viscosity, chip moisture, and SO_2 levels - structured by operator and time slot. This data, entered consistently by plant operators across shifts, formed the manual layer of the mill's unified data foundation. When operators saw their inputs surfacing in process improvement and optimization, the discipline of structured entry became self-reinforcing.

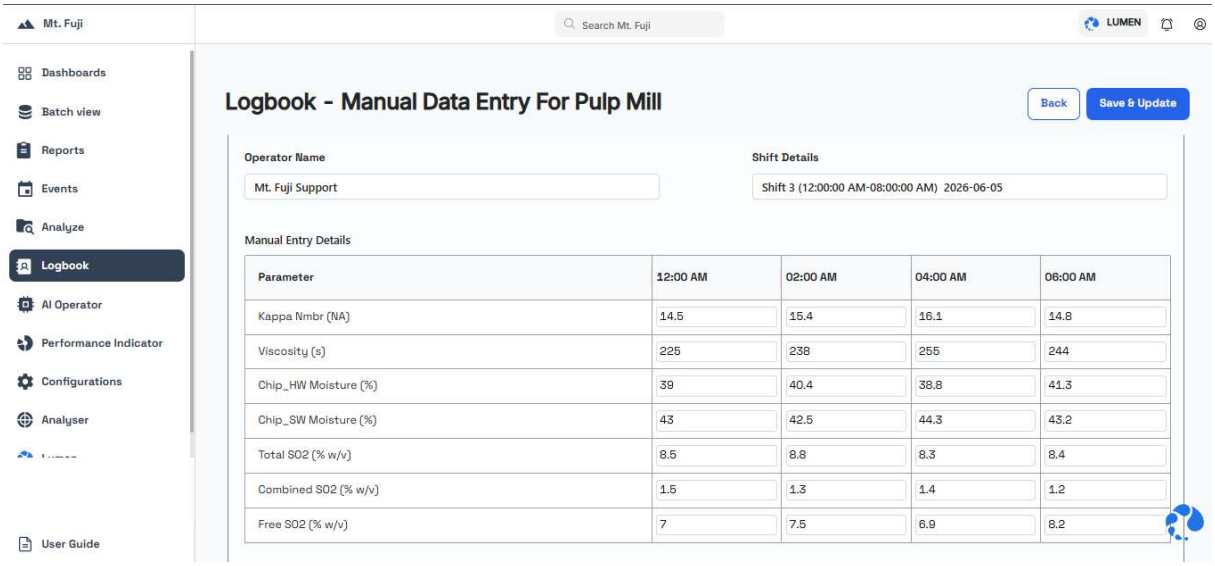

Figure 3: Mt. Fuji E-Logbook - shift-wise manual entry of critical process parameters by plant operators

4. Case Study: Golden Batch Analysis

4.1 Background and Pilot Scope

With a structured data foundation in place, the full analytical potential of historical process data becomes accessible. The Golden Batch Analysis, conducted at a pulp mill, demonstrated this concretely. The mill operated a sulphite (Acidic) pulping process with six batch digesters. The pilot addressed a specific operational question: what process conditions consistently produced pulp that simultaneously met both Kappa (14–16) and Viscosity (220–260) s quality targets?

The pilot scope covered three deliverables: identification of golden batches - those simultaneously achieving both target ranges - from the historical dataset; definition of the operating parameter ranges associated with golden batches, expressed as a per-digester recipe; and recommendations for improving golden batch consistency.

The analysis leveraged a dataset of 4,000 batch records, integrated with continuous process signals from the chemical preparation system and downstream washing sections into a unified process knowledge graph. This structured foundation provided comprehensive coverage of process variability, enabling the identification of operating conditions most consistently associated with successful outcomes and the subsequent definition of a golden operating envelope.

4.2 Principal Finding: Limited Golden Batch Achievement

The central finding was clear: under stable operating conditions, only 23% of batches simultaneously achieved both Kappa and Viscosity targets.

A decomposition of performance revealed where the primary constraint lay. Kappa control was relatively effective, with 72.3% of batches within target, indicating that delignification was generally well managed. In contrast, Viscosity performance was significantly more variable, with only 26.5% of batches meeting the desired range. The principal limitation was therefore not Kappa control, but the ability to consistently preserve pulp strength.

The mill data exhibited several distinct sources of variation, and separating them was central to interpreting the results. Batch-to-batch variation was the most prominent: Kappa values were relatively tightly distributed around the target band, whereas Viscosity showed a markedly wider spread, with a substantial tail of batches falling below the lower specification limit. A second source was input-driven variation, most clearly the influence of chip moisture, where golden batch rates differed by more than ten percentage points between the driest and wettest quartiles. A third was operational variation across shifts and over the eight-month window, reflecting differences in chemical dosing, cook timing, and chip feedstock as production conditions changed. Time-At-Max, the primary operator-controlled cooking severity parameter, also exhibited significant variation both batch-to-batch and across digesters, highlighting differences in operating practices used to achieve target quality.

By contrast, vessel-to-vessel variation was comparatively small (a 20.1%–23.4% range across the six digesters), confirming that the dominant variation was driven by process recipe and input conditions rather than by individual equipment behaviour. Characterising these variations separately was only possible because each measurement carried its process context, and it is this decomposition that translated raw scatter into actionable levers.

The implication was clear: improving Viscosity consistency represented the single largest opportunity to increase the overall golden batch rate.

An important structural observation reinforced this conclusion. Golden batch performance varied only marginally across digesters (20.1%–23.4%), indicating that no single vessel was underperforming. The problem was systemic, rooted in the process recipe rather than digester-specific characteristics. Improvements to the operating strategy were therefore expected to benefit the entire mill.

This pattern is illustrated in Figure 4, which plots the joint distribution of Kappa Number and Viscosity for all batches against the two target windows (Kappa 14–16 and Viscosity 220–260 s). The scatter shows the bulk of batches clustering within the Kappa target band while spreading well beyond the Viscosity band, so that only the batches falling inside both shaded ranges (the golden batches) account for the 23% joint success rate. The figure makes the bottleneck visually apparent: the horizontal (Viscosity) spread, rather than the vertical (Kappa) spread, is what pushes most batches outside the combined target zone.

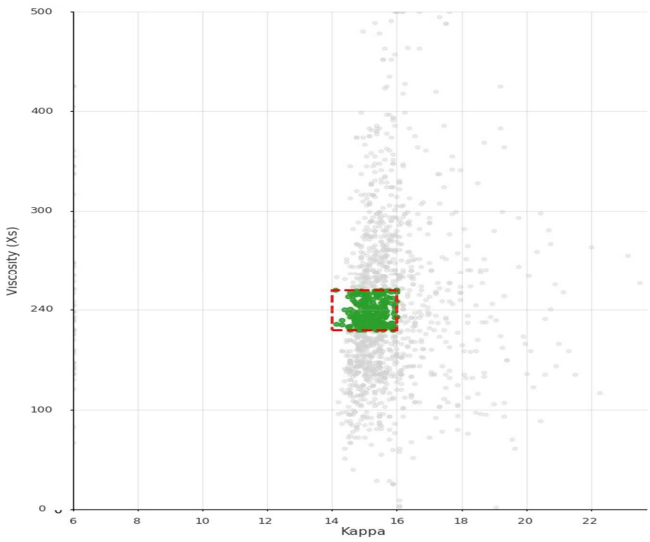

Figure 4: Joint distribution of Kappa Number and Viscosity across all batches

4.3 The Effect of Chip Moisture on Batch Quality Outcomes

One of the most practically significant findings concerned chip moisture.

Conventional mill logic often treats higher moisture content as undesirable due to its impact on cooking efficiency. The analysis, however, revealed the opposite relationship: higher chip moisture significantly improved golden batch performance, increasing success rates from 14.2% in the driest chips to 25.6% in the wettest chips. As summarised in Table 1, the golden batch rate rose monotonically across the four moisture quartiles, climbing from 14.2% (Q1, driest) through 17.5% and 20.1% to 25.6% (Q4, wettest), a near-doubling that establishes chip moisture as one of the clearest single levers identified in the dataset.

The process mechanism is likely straightforward: wetter chips allow more uniform acid penetration during impregnation, reducing the formation of localised overcooked regions that depress both Kappa and Viscosity performance.[3, 6] This finding inverted a common operating assumption - rather than minimising chip moisture, the analysis recommended establishing a minimum moisture threshold of 38% and avoiding the processing of excessively dry chips.

Table 1. Golden batch rate by chip moisture quartile.

Total Moisture Quartile	Moisture Range (%)	Golden Batch Rate
Q1 (driest)	Below 37.5%	14.2%
Q2	37.5% – 39.5%	17.5%
Q3	39.5% – 41.1%	20.1%
Q4 (wettest)	Above 41.1%	25.6%

4.4 Downstream Process Evaluation

The analysis further extended its investigative scope to downstream quality metrics and process variables, encompassing brightness, whiteness, conductivity, and filtrate quality. These parameters were scrutinised to identify potential causal linkages between digester outcomes and subsequent unit operations.

While the primary drivers of quality were concentrated within the cooking phase, this integration of batch-specific and continuous process data across operations proved invaluable. Establishing this cross-functional data lineage provides the necessary infrastructure for future process-wide optimisation initiatives.

A significant opportunity exists in extending this batch-level traceability into bleaching stages. By establishing rigorous links between digester performance, bleachability, and chemical demand, mills can begin to quantify the downstream economic value of cooking consistency and drive targeted bleach chemical optimisation.

These findings underscore the broader imperative for mills to bridge the data silos between cooking, washing, and bleaching. Such data connectivity enables a comprehensive understanding of how upstream operational decisions dictate downstream performance, resource efficiency, and final product quality.

4.5 The Golden Batch Operating Envelope

The central analytical deliverable of the study was the definition of a golden operating envelope: the range of process conditions most consistently associated with successful batches.

Across historical golden batches, the operating envelope was characterized by a balanced SO_2-to-wood ratio, controlled free and total SO_2 levels, optimized net acid dosage, moderate total cook duration, reduced time at maximum temperature, and consistently higher chip moisture.

Together, these parameters define a practical and actionable operating recipe. Maintaining process conditions within this envelope provides the clearest pathway toward improved consistency and higher golden batch achievement.

4.6 Vessel-Level Analysis

A key structural insight emerged from comparing golden batches across all digesters. While chemical recipes were highly consistent across vessels, cook severity varied significantly.

Some digesters consistently achieved golden conditions using a shorter and less severe cooking profile, while still meeting both Kappa and Viscosity targets. This suggests that multiple operational pathways can achieve the desired quality outcome - but that lower-severity cooking may offer a substantial advantage for Viscosity preservation.

This finding represented a major optimisation opportunity. If the lower-severity cooking strategy could be replicated across all digesters while maintaining Kappa control through chemistry adjustments, the result would be a significant increase in overall golden batch consistency

5. Implications for the Indian Paper Industry

5.1 A Replicable Methodology

The methodology described in this paper - data foundation first, analytical insight second - is not specific to sulphite pulping or any particular mill configuration. It is a replicable sequence that any mill can follow, adapted to its own process, instrumentation level, and quality KPIs. The starting point is always the same: an honest assessment of current data capture practices, identification of the critical variables that influence the quality outcomes that matter, and a disciplined programme to capture those variables with consistency and context.

5.2 Response to Common Implementation Concerns

‘We don’t have the data infrastructure.’ The Golden Batch Analysis was conducted on data from improved manual and laboratory measurement routines, without new sensor deployment. The infrastructure required is a digital data capture tool, a historian or structured database, and consistent recording discipline. This is accessible to any mill.

‘Our lab results and historian data are in completely different systems - can they actually be joined?’ Every mill that has undergone this programme has presented the same starting condition. The critical enabler is not system consolidation but the batch identifier: if a laboratory result can be linked to the batch it was drawn from, and that batch has a timestamp in the historian, the join is possible.

‘Our process is too complex / too variable for this approach.’ Process complexity is the argument for better data, not against it. The more variable the process, the more important it is to understand what conditions are associated with good outcomes. The analysis found consistent patterns - the golden envelope - across six different digesters and eight months of variable production. Complexity is navigable with good data. It is not navigable without it.

5.3 Convergent Benefits of the Data Foundation Approach

For the Indian paper industry, the data foundation approach addresses multiple pressures simultaneously. Better quality data enables chemical optimisation, reducing input costs. It enables consistency improvement, reducing customer complaints and reject rates. It enables sustainability optimization - lower effluent loads, reduced water consumption, lower energy intensity - because the variables that drive these outcomes become visible and controllable.

Most importantly, it positions mills for the AI-driven optimisation tools that are becoming standard in the global paper industry. A mill that builds its data foundation today is one to two years ahead of a mill that defers this work and then discovers, when deploying an AI system, that the data it holds is insufficient to train a reliable model.

The structured data layer - when organised as a process knowledge graph - also becomes the foundation for the next generation of AI tools: systems that can answer complex operational questions by navigating connected process context, rather than searching through disconnected records.

6. Conclusion

This paper presented a two-part argument grounded in practical case study evidence. First, data foundation building, the systematic capture and contextualisation of process data, is the necessary first step for AI in pulp and paper mills: the limiting factor in most mills is not the availability of sophisticated algorithms but the quality, continuity, and contextual richness of the underlying data. Second, once that foundation was in place, the Golden Batch Analysis delivered precise, evidence-based operational insight, identifying that only 23% of digester batches simultaneously achieve both Kappa and Viscosity targets, with viscosity the primary bottleneck, cook duration the primary lever, chip moisture a systematically underutilised input, and the acid-chemistry recipe portable across all six digesters while cook severity required per-vessel calibration. Because this descriptive approach is replicable across the Indian paper industry without large capital investment in sensors, and because the same data foundation can subsequently support predictive and prescriptive applications, it represents an immediate and practical pathway to competitive advantage for an industry facing simultaneous pressure on cost, quality, and sustainability.

References

1. Jiang, H. and Li, G. (2021). Application of machine learning in paper and pulp industry: A review. TAPPI Journal, 20(4), 237–252.
2. Casey, J.P. (1980). Pulp and Paper: Chemistry and Chemical Technology, 3rd ed. Wiley-Interscience, New York.
3. Gullichsen, J. and Fogelholm, C.-J. (1999). Chemical Pulping. Fapet Oy, Helsinki.
4. Smook, G.A. (2016). Handbook for Pulp and Paper Technologists, 4th ed. TAPPI Press, Atlanta.
5. Indian Paper Manufacturers Association (IPMA). (2024). Annual Report 2023–24: Indian Paper Industry Overview. IPMA, New Delhi.
6. Lindstrom, M.E. and Gellerstedt, G. (2008). Sulphite pulping. In: Henriksson, G. (ed.) Wood Chemistry and Biotechnology. De Gruyter, Berlin, pp. 201–228.