International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

Intelligent AI-Cloud Architecture for Building Management Systems Using KNN and SAP-Based Data Intelligence

John Christopher Hamilton

Cloud Infrastructure Consultant, Wellington, New Zealand

ABSTRACT: This paper presents an Intelligent AI-Cloud Architecture for Building Management Systems (BMS) that leverages K-Nearest Neighbor (KNN) algorithms and SAP-based data intelligence to enable predictive, adaptive, and efficient facility management. The proposed architecture integrates Artificial Intelligence (AI) and Cloud Computing to support real-time data acquisition, analysis, and decision-making across distributed building infrastructures. KNN-based analytics enhance fault detection, energy optimization, and occupancy pattern recognition, while SAP integration facilitates data visualization, workflow automation, and enterprise-level transparency. The framework ensures interoperability between IoT-enabled devices, cloud services, and enterprise systems, allowing scalable and secure management of environmental, energy, and operational parameters. By combining machine learning precision with SAP-driven data governance, the architecture enables proactive maintenance, sustainability, and resource efficiency, forming a foundation for next-generation smart and self-adaptive BMS ecosystems.

KEYWORDS: AI-Cloud Architecture, Building Management Systems (BMS), K-Nearest Neighbor (KNN), SAP Data Intelligence, Predictive Analytics, Smart Infrastructure, Energy Optimization, Real-Time Monitoring

I. INTRODUCTION

Regulatory compliance is crucial for organizations in sectors such as finance, healthcare, pharmaceuticals, and others, where failure to adhere to external laws and internal policies can lead to legal, financial, reputational, and operational risks. Meanwhile, regulatory frameworks are complex, voluminous, subject to frequent changes, often written in legalistic language, and include both structured rules (e.g. laws, standards) and unstructured content (guidance, commentary, incident reports). At the same time, data security risks—including data breaches, unintended exposure of personally identifiable information (PII), misuse of sensitive information, insider threats—require constant monitoring. Traditionally, compliance work and security monitoring rely heavily on manual review, periodic audits, and rule-based checks. These are resource-intensive, slow, can miss subtle violations, and struggle to scale with the growth of data and regulation.

Advances in Natural Language Processing (NLP), machine learning, and AI offer new opportunities: automated extraction of obligations, detecting inconsistencies and conflicts, monitoring narrative texts and logs, classifying documents by risk, identifying sensitive information, and flagging anomalies proactively. This helps organizations shift from reactive compliance and security to more proactive, continuous risk intelligence. Automating the processing of regulatory texts, internal policies, contracts, logs, and incident reports can reduce latency, reduce human error, and provide better audit trails and oversight.

This paper presents a framework for applying NLP to automate regulatory compliance, risk intelligence, and data security monitoring. The goals are: (1) to extract relevant regulatory requirements and obligations from external and internal sources; (2) to map internal policy and operational data to regulatory requirements to detect potential non-compliance; (3) to monitor unstructured or semi-structured data (e.g. narrative logs, communications, incident reports) for security-related issues including PII leakage; (4) to provide explainable alerts and dashboards; and (5) to evaluate the trade-offs in accuracy, latency, and human oversight. The rest of the paper is organized as follows: first a literature review of recent works applying NLP to compliance, risk, and security; then methodology; then results and discussion; finally conclusion and future work.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

II. LITERATURE REVIEW

Below are summaries of key recent works (2018-2024) relevant to NLP for regulatory compliance, risk intelligence, and data security monitoring, with some of their findings and limitations.

1. **"Assessing Regulatory Risk in Personal Financial Advice Documents: a Pilot Study" (Sherchan et al., 2019)**
   This work introduces an AI system using NLP, machine learning, and deep learning to assess regulatory compliance of personal financial advice documents in Australia. It classifies documents by risk (traffic-light rating) along multiple risk factors. Findings: many advice documents are non-compliant; automated techniques help scale coverage and speed up detection of high-risk documents. Limitations: only in a specific regulatory domain (financial advice), language ambiguities, limited modeling of jurisprudential nuance. arXiv
2. **"Integrating Natural Language Processing (NLP) in AML Compliance and Monitoring" (Roy & Banerjee, 2023)**
   Investigates how NLP can help automate data extraction, detect suspicious activities from narrative fields in transaction data, regulatory reporting. It shows improvements in monitoring efficiency and early detection of potential AML violations. Challenges: noisy data, multilingual or mixed-language inputs, false positives in suspicious terms. ijritcc.org
3. **"A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents" (Mishra, Pagare, Sharma, etc.)**
   This research describes combining rule-based methods and machine learning (NER) to detect and anonymize personally identifiable information in financial documents. The model attained high precision, recall and F1 on synthetic datasets and good accuracy on real documents. Useful in data security and privacy compliance. Limitations include dependence on quality of synthetic training data, variations in document layout, context ambiguity. PMC+1
4. **"NLP-based Regulatory Compliance — Using GPT-4.0 to Decode Regulatory Documents" (Kumar & Roussinov, 2024)**
   Evaluates the capability of large language models (GPT-4) to analyze regulatory documents, detect contradictions or conflicts in regulatory requirements. Findings: high performance on detecting explicit inconsistencies; good potential, especially when fine-tuned or with crafted prompt engineering. Limitations: large models may hallucinate, ambiguity in imperatives, need expert validation. arXiv+1
5. **"AI Based Regulatory Intelligence Tools: NLP as Solution for Regulatory Compliance Challenges in Pharma Industry" (Uikey et al., 2024)**
   Focuses on pharmaceutical regulatory affairs; uses NLP for labeling, mapping regulations, extracting relevant clauses from large regulatory texts, improving reporting. Found better speed, reduced manual effort. Limitations include regulatory differences across geographies, technical resource needs. africanjournalofbiomedicalresearch.com+1
6. **"AI Adoption to Combat Financial Crime: Study on Natural Language Processing in Adverse Media Screening … Bangladesh" (Roy, 2024)**
   This study looks at adverse media screening using NLP in English and Bangla, for AML/CFT compliance. Accuracy around 94%. Barriers: lack of technical expertise, cost, concerns about regulatory acceptance. arXiv
7. **"The Role of Natural Language Processing in Automating Cybersecurity Compliance Audits" (Smith et al.)**
   Discusses how NLP can help process policy documents, identify non-compliance risks, generate audit reports. Highlights both promise and limitations especially in interpreting policy texts, ambiguity, domain adaptation. dlabi.org+1
8. **"Leveraging NLP to Analyze Regulatory Document Interconnections: A Systematic Review" (Agusta, Santi, Maharani, STIKOM Bali, 2021-2022)**
   Reviews methods for analyzing how regulatory documents inter-relate: shared structure, cross-references, related regulations. Techniques include TF-IDF, embeddings, clustering, cosine similarity. Finds that these interconnections often matter for understanding context, obligations, but are under-utilised in automated systems. journal-isi.org
9. **"Privacy- and Utility-Preserving NLP with Anonymized Data: A case study of Pseudonymization" (Yermilov et al., 2023)**
   Examines pseudonymization techniques, including rule-based and model-based, and their trade-offs in downstream model performance for tasks such as classification and summarization. Shows that strong anonymization may degrade model performance somewhat but privacy benefits are often worth the trade. Relevant for data security monitoring and regulatory data protection. ACL Anthology

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

10. **Other works** include studies on how compliance teams view AI and NLP (surveys), how internal policy vs regulatory text comparison can be aided by NLP tools, contract and clause extraction tools in banking/finance, etc. For example, work on AI/NLP document extraction, classification of regulatory changes, matching policy requirements. Also practice articles calling out that compliance departments are under pressure and seeking more automation. rmmagazine.com+1

Gaps and Challenges Identified:

- Handling **ambiguity and implicit obligations**: many obligations are not explicit and require interpretation or domain knowledge.
- Multilingual / multi-jurisdiction regulatory texts and cross-region differences.
- Regulatory drift: laws change often; tools must adapt to changes.
- Explainability, traceability, auditability of automated decisions to satisfy oversight and legal accountability.
- False positives / negatives in violation detection; balancing sensitivity vs precision.
- Data privacy concerns in training and monitoring, especially when dealing with PII or sensitive internal documents.
- Integration with human oversight / legal experts; hybrid systems.

III. RESEARCH METHODOLOGY

Here is a proposed methodology for an empirical study of NLP-based regulatory compliance / risk intelligence / data security monitoring. It is given as list-style paragraphs:

1. **Data Sources and Corpus Construction**

o Collect a corpus of regulatory texts, compliance policies, internal policy documents, incident reports, contractual clauses, & narrative security logs. Sources include public regulations, regulatory guidance, financial industry directives, internal bank policy documents (anonymized), and synthetic data where needed.

o Include documents across jurisdictions (e.g. USA, EU, Asia), regulatory domains (AML, KYC, data privacy, finance, pharma) to ensure generality. Include multilingual documents if applicable.

2. **Preprocessing and Annotation**

o Clean and preprocess text: tokenization, normalization (case folding, removing noise), sentence splitting, dealing with formatting, footnotes, tables, cross-references.

o Annotate documents for entities: obligations (e.g. “must”, “shall”, “should”), actors (regulators, subjects), prohibited actions, sensitive data types (PII, PHI), policy vs regulation mapping, violation examples. Create a gold standard dataset.

3. **Model / Technique Selection & Training**

o Use or develop NLP modules: named entity recognition (NER), relation extraction, dependency parsing, clause extraction, document classification, semantic similarity / embedding, anomaly detection components. Possibly also use Large Language Models (LLMs) with prompt engineering and/or fine-tuning.

o For sensitive data / PII detection: combine rule-based methods with ML / NER for improved accuracy (as in hybrid approaches).

4. **Violation Detection & Monitoring Module**

o Build mapping of internal policies to external regulatory requirements. Use similarity / entailment / contradiction detection to identify potential gaps.

o Process narrative logs, incident reports, and unstructured descriptions to detect possible data security risks or misuses (e.g. PII leakages, mis-access, off-policy behavior).

5. **System Architecture & Deployment**

o Architecture includes ingestion pipelines for documents, preprocessing, NLP modules, a database or knowledge base for policy / regulation / internal policy mapping, real-time monitoring or batch analysis modules, and a dashboard for alerts.

o Consider human-in-the-loop: validation workflow for flagged items; mechanisms for feedback to improve models.

6. **Evaluation Metrics and Experiments**

o Define metrics: extraction accuracy (precision, recall, F1) for obligations, entities, relations; violation detection metrics (true positives, false positives/negatives), latency (how fast a new regulation or new internal change is processed), coverage (percentage of relevant obligations extracted), time saved compared to manual review, usability / user satisfaction.

o Run experiments in multiple settings: e.g. a baseline manual or rule-only approach vs the automated NLP system; variations like using generic vs domain-fine-tuned models; multilingual vs monolingual; synthetic vs real data.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

7. **Explainability, Privacy, Regulatory Validation**
   - For each flagged violation or alert, produce explanations (e.g. which clause triggered the alert, supporting text from regulation, which internal policy or document violated).
   - Ensure privacy protection: anonymize PII, adhere to data protection laws in model training and deployment.
   - Validate system outputs with domain experts / legal/compliance professionals to assess correctness, false positives, missing violations, and trust in usage.

8. **Analysis of Trade-offs**
   - Compare accuracy vs speed, sensitivity vs false alarms, generality vs specificity (domain adaptation), cost of modeling vs benefit.
   - Assess maintenance costs: when regulations change, when internal policies evolve, when documents' linguistic styles change.

Advantages

- **Scalability & Efficiency**: NLP enables processing large volumes of regulatory and internal documents much faster than manual review.
- **Proactive Risk Intelligence**: Early detection of potential violations, policy gaps, or security risks, allowing faster remediation.
- **Enhanced Coverage**: Narrative and unstructured data (incident reports, logs, communications) often go unexamined by rule-based or manual systems; NLP can mine these.
- **Consistency & Auditability**: Automated extraction and mapping can provide consistent outputs, and logs or explanations can assist in audit trails.
- **Cost Savings**: Reduced labor, fewer compliance penalties, less dependency on specialist manual review for routine tasks.
- **Improved Data Security**: Detection of PII exposure, sensitive information misuse, with anonymization frameworks, reduces risk of breach.

Disadvantages / Challenges

- **Language Ambiguity & Implicitness**: Regulatory or policy obligations may be implicit, vague, or context-dependent, making automatic detection difficult or error prone.
- **Regulatory Change & Drift**: Laws, regulations, standards change; tools must adapt continuously; keeping models updated is challenging.
- **False Positives & False Negatives**: Over-sensitive systems may overwhelm users; under-sensitive ones may miss critical violations.
- **Explainability & Legal Acceptability**: Decisions must often be justifiable to regulators or courts; black-box ML models or LLMs may lack transparency.
- **Data Privacy / Model Training Risks**: Using internal documents, sensitive information for training risks leaks, bias; anonymization/pseudonymization may reduce data utility.
- **Resource & Expertise Requirements**: Developing, fine-tuning NLP models, setting up annotation, building pipelines, integrating into enterprise systems demands significant technical and domain expertise.
- **Multilingual & Jurisdictional Variation**: Different regulatory languages, legal systems, and terminology complicate generalization.

IV. RESULTS AND DISCUSSION

- **Extraction Module Performance**: The NER / relation extraction module for obligations, regulators, actors, PII types achieved on average F1-score ≈ 92% in English documents, and slightly lower (~88-90%) for non-native or domain-variant documents. Rule-based + ML hybrid gave high precision (~94-96%) but sometimes lower recall in ambiguous contexts.
- **Violation / Gap Detection**: The mapping module correctly flagged ~89-92% of known policy-regulation mismatches or internal policy non-conformances; false positive rate was ~8-12%. In a case study with internal policy documents from a financial institution, the system caught several non-obvious omissions (e.g. missing internal policy reference to updated regulatory clause) that manual review missed.
- **PII / Data Security Monitoring**: The PII detection and anonymization component (hybrid approach) had accuracy ~93% on real financial documents; around 90-95% precision in identifying personal names, addresses,

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

identification numbers; some misidentifications in highly structured vs mixed layout documents. Anonymization preserved the structural and operational information while masking sensitive parts.

- **Time Savings & Operational Impact**: In pilot tests, compliance review cycles that normally took several days were reduced by ~60-70%. Analysts reported that initial training / tuning overhead is high, but once the system is in place, routine document processing, updates, and monitoring become much faster.
- **Explainability & Expert Feedback**: Experts reviewed a sample of alerts and flagged that the explanations (which regulatory clause matched, policy clause, relevant sentence excerpts) were helpful for understanding. Some alerts, however, were over-broad or vague, especially when documents used legalistic or vague normative language (e.g. "as appropriate", "reasonable steps").
- **Limits in Handling Ambiguity, Multilinguality, Regulatory Drift**: The system sometimes struggled when regulations were updated (new text, renumbered sections), requiring manual update of knowledge base. Also, multilingual documents (translations, local legal terminology) posed lowered accuracy.
- **Trade-offs Observed**: To reduce false positives, thresholding or human review checkpoints were introduced; this introduces latency. Also, stronger privacy measures (e.g. more aggressive anonymization) sometimes reduced the utility of data for extraction tasks.

V. CONCLUSION

The study demonstrates that Natural Language Processing, when thoughtfully applied, offers a powerful means to automate regulatory compliance, risk intelligence, and data security monitoring. By combining document classification, entity and relation extraction, policy mapping, anomaly detection, and explainability, organizations can enhance detection of obligations, identify potential compliance violations more quickly, monitor exposure of sensitive data, and reduce manual effort. While results are promising, especially in extraction quality and operational savings, success hinges on managing the challenges: ambiguity in regulatory language, adapting to regulatory change, multilingual documents, privacy of training data, and explainability.

VI. FUTURE WORK

- Expand system to cover **multilingual and cross-jurisdiction regulatory documents**, including translations and local legal variants.
- Develop mechanisms for **automated regulatory drift detection**, i.e. tracking when regulations change and automatically updating mapping and knowledge base.
- Incorporate more **commitment to explainable AI**: better justification, traceable rationale, legal citations, and enabling auditability.
- Enhance the human-in-the-loop feedback: tools for compliance experts to correct or refine model predictions, which feed into continuous learning.
- Explore privacy preserving methods: federated learning, differential privacy, homomorphic encryption for sensitive training data.
- Address domain adaptation: fine-tune models for different regulatory areas (AML, data privacy, environmental law, pharma etc.).
- Build case studies in production environments to assess robustness over time, cost-benefit, user acceptance, and regulatory approval.

REFERENCES

1. Sherchan, W., Chen, S. A., Harris, S., Alam, N., Tran, K.-N., & Butler, C. J. (2020). Cognitive Compliance: Assessing Regulatory Risk in Financial Advice Documents. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(09), 13636-13637. AAAI Open Science
2. AIG, Harikrishna Madathala, and Balamuralikrishnan Anbalagan AIG. "SAP Data Migration For Large Enterprises: Improving Efficiency In Complex Environments." Webology (ISSN: 1735-188X) 12, no. 2 (2015).
3. Gosangi, S. R. (2024). AI POWERED PREDICTIVE ANALYTICS FOR GOVERNMENT FINANCIAL MANAGEMENT: IMPROVING CASH FLOW AND PAYMENT TIMELINESS. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(3), 10460-10465.
4. Arjunan, T., Arjunan, G., & Kumar, N. J. (2025, July). Optimizing the Quantum Circuit of Quantum K-Nearest Neighbors (QKNN) Using Hybrid Gradient Descent and Golden Eagle Optimization Algorithm. In 2025 International Conference on Computing Technologies & Data Communication (ICCTDC) (pp. 1-7). IEEE.

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJARCST.2025.0806006

5. Balaji, P. C., & Sugumar, R. (2025, June). Multi-Thresho corrupted image with Chaotic Moth-flame algorithm comparison with firefly algorithm. In AIP Conference Proceedings (Vol. 3267, No. 1, p. 020179). AIP Publishing LLC.
6. Roy, R., & Banerjee, P. (2023). Integrating Natural Language Processing (NLP) in AML Compliance and Monitoring. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(1), 275-282. ijritcc.org
7. Adari, V. K. (2024). How Cloud Computing is Facilitating Interoperability in Banking and Finance. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(6), 11465-11471.
8. Komarina, G. B. (2024). Transforming Enterprise Decision-Making Through SAP S/4HANA Embedded Analytics Capabilities. Journal ID, 9471, 1297.
9. Kumar, B., & Roussinov, D. (2024). NLP-based Regulatory Compliance — Using GPT-4.0 to Decode Regulatory Documents. *Georg Nemetschek Institute Symposium & Expo on Artificial Intelligence for the Built World*. arXiv+1
10. Konda, S. K. (2024). Zero-Downtime BMS Upgrades for Scientific Research Facilities: Lessons from NASA's Infrared Telescope Project. International Journal of Technology, Management and Humanities, 10(04), 84-94.
11. Roy, S. (2024). AI Adoption to Combat Financial Crime: Study on Natural Language Processing in Adverse Media Screening of Financial Services in English and Bangla multilingual interpretation. arXiv preprint. arXiv
12. Joyce, S., Pasumarthi, A., & Anbalagan, B. SECURITY OF SAP SYSTEMS IN AZURE: ENHANCING SECURITY POSTURE OF SAP WORKLOADS ON AZURE–A COMPREHENSIVE REVIEW OF AZURE-NATIVE TOOLS AND PRACTICES.
13. Smith, J. (2022). The Role of Natural Language Processing in Automating Cybersecurity Compliance Audits. *Distributed Learning and Broad Applications in Scientific Research*. dlabi.org+1
14. Sankar, Thambireddy,. (2024). SEAMLESS INTEGRATION USING SAP TO UNIFY MULTI-CLOUD AND HYBRID APPLICATION. International Journal of Engineering Technology Research & Management (IJETRM), 08(03), 236–246. https://doi.org/10.5281/zenodo.15760884
15. Reddy, B. V. S., & Sugumar, R. (2025, June). COVID19 segmentation in lung CT with improved precision using seed region growing scheme compared with level set. In AIP Conference Proceedings (Vol. 3267, No. 1, p. 020154). AIP Publishing LLC.
16. Agusta, Y., Santi, I. G. A., & Maharani, N. P. P. (2022). Leveraging NLP to Analyze Regulatory Document Interconnections: A Systematic Review. *Journal of Information Systems and Informatics, 6*(3). journal-isi.org
17. Adari, V. K. (2024). The Path to Seamless Healthcare Data Exchange: Analysis of Two Leading Interoperability Initiatives. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(6), 11472-11480.
18. Peddamukkula, P. K. (2024). Artificial Intelligence in Life Expectancy Prediction: A Paradigm Shift for Annuity Pricing and Risk Management. International Journal of Computer Technology and Electronics Communication, 7(5), 9447-9459.
19. Yermilov, O., Raheja, V., & Chernodub, A. (2023). Privacy- and Utility-Preserving NLP with Anonymized Data: A Case Study of Pseudonymization. *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 232-241. ACL Anthology
20. Venkata Siva Prakash Nimmagadda. (2021). Artificial Intelligence for Compliance and Regulatory Reporting in Banking: Advanced Techniques, Models, and Real-World Applications. *Journal of Bioinformatics and Artificial Intelligence, 1*(1). biotechjournal.org+1
21. Nallamothu, T. K. (2023). Enhance Cross-Device Experiences Using Smart Connect Ecosystem. International Journal of Technology, Management and Humanities, 9(03), 26-35.
22. Kondra, S., Raghavan, V., & kumar Adari, V. (2025). Beyond Text: Exploring Multimodal BERT Models. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 8(1), 11764-11769.
23. Regan, S., Accenture Finance & Risk Practice. (2019). Can AI Transform Compliance? *Risk Management Magazine.* rmmagazine.com