Journal of Integrated Science, Technology and Management, 2025; 1(2):15-21

Journal of Integrated Science, Technology and Management

Research Article

Automated Data-Quality Assessment and Intelligent Feature Engineering for Large-Scale Predictive Analytics

Anvar Tursunov[1*], Mohd Asif Khan[2]

[1]*Department of History, Termez University of Economics and Service, Termez, Uzbekistan*
[2]*Researcher*

ARTICLE INFO

Article history:
Received: 25-05-2025
Revised: 11-07-2025
Accepted: 19-09-2025
Published: 05-11-2025

Keywords:
Data quality assessment, automated feature engineering, predictive analytics, data profiling, concept drift, machine learning, scalable data pipelines, explainability.

DOI: 10.66592/nfw294bd

ABSTRACT

Reliable predictive analytics depends not only on the capacity of learning algorithms but also on the quality, semantic consistency, and predictive usefulness of input data. This paper develops a conceptual, modular framework that combines automated data-quality assessment with intelligent feature engineering for large-scale tabular analytics. The proposed pipeline profiles completeness, validity, uniqueness, consistency, timeliness, and distributional stability; applies risk-aware remediation; constructs statistically and semantically meaningful features; and uses leakage-safe validation to select transformations. A multi-objective optimization formulation balances predictive utility, computational cost, feature stability, and quality risk. An implementation protocol for distributed processing, experiment tracking, and monitoring is specified. To demonstrate reporting and interpretation, a transparent illustrative numerical scenario compares four pipeline configurations across accuracy, F1 score, AUROC, and processing time. These figures are hypothetical examples, not empirical findings. The framework contributes an auditable design that links data-quality decisions to downstream model reliability and provides testable hypotheses for future benchmarking on real datasets.

INTRODUCTION

Large-scale predictive analytics has become central to decision-making in healthcare, manufacturing, finance, transportation, education, and public administration. Organizations routinely combine transactional records, sensor streams, application logs, and externally sourced datasets to estimate outcomes and allocate resources. Yet the value of a predictive model is constrained by the reliability of the data entering its training and inference pipelines. A sophisticated learner may fit patterns in corrupted records, exploit accidental identifiers, or produce unstable predictions when operational conditions change. Consequently, data preparation should be treated as a governed modeling activity rather than an informal preliminary step.

Data quality is multidimensional. Completeness describes the availability of expected values; validity measures conformity to declared rules; uniqueness concerns duplicate entities; consistency addresses conflicting representations; timeliness captures the age and freshness of information; and accuracy refers to agreement with the underlying real-world state. These dimensions cannot always be reduced to a single universal score. A missing field can be harmless for one task and decisive for another. A record that is syntactically valid can still be semantically implausible. Effective assessment must therefore connect technical checks to domain context, target definition, and the consequences of incorrect predictions.

At scale, manual inspection becomes impractical. Data may arrive from hundreds of services with different schemas, update frequencies, and ownership structures. A static rule library can detect known violations, but new anomalies emerge as business processes evolve. Automated profiling can identify changing distributions, unexpected categories, abnormal missingness, and shifts in relationships between variables. Such detection is particularly valuable when a dataset contains millions of rows and thousands of candidate columns. Nevertheless,

***Corresponding Author:** Anvar Tursunov
Address: Department of History, Termez University of Economics and Service, Termez, Uzbekistan
Email ✉ anvar_tursunov@tues.uz **ORCID:** 0009-00006-2981-4376

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

© 2025, Anvar Tursunov and Mohd Asif Khan, This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

automation creates a governance problem: aggressive cleaning can silently remove minority populations, distort rare events, and erase information relevant to the outcome.

Missing-data handling illustrates this tension. Simple mean or median substitution is inexpensive, but it can reduce variance and obscure conditional relationships. More complex approaches can exploit correlations, yet they risk introducing artificial patterns and increasing processing cost. Missingness may itself convey information, such as a skipped measurement resulting from an operational policy. A reliable workflow should first examine missingness mechanisms and downstream sensitivity, then choose an imputation policy appropriate to each feature and deployment context. Critically, imputation statistics must be learned only from training partitions to prevent information leakage.

Outlier treatment requires comparable care. Extreme values may represent recording errors, sensor faults, rare but legitimate cases, or the very events a model is intended to detect. Global thresholding can be misleading in heterogeneous populations. Context-aware detection, robust statistics, and isolation-based methods offer alternatives, but each must be assessed against the target task. In fraud detection, for example, removing unusual transactions can degrade recall. Thus, the system should preserve traceable decisions and distinguish a quality alert from a mandatory deletion instruction.

Feature engineering transforms raw observations into representations that expose relationships more effectively to learning algorithms. Useful transformations include categorical encoding, temporal aggregation, ratios, interaction terms, rolling statistics, and domain-derived indicators. Manual feature construction often produces strong results but demands expertise and repeated experimentation. Automated feature engineering explores transformation candidates systematically, potentially discovering useful representations that analysts overlook. However, uncontrolled generation causes combinatorial growth, multicollinearity, high memory use, and increased leakage risk.

The interaction between data quality and feature engineering is especially important. A ratio may amplify measurement noise when its denominator is near zero; temporal aggregates may hide missing observations; categorical encoders may fail on new labels; and complex interactions may magnify subgroup bias. A quality-aware generator should attach provenance and reliability metadata to candidate features and reject transformations that violate temporal availability or semantic constraints. The selection objective should reward generalizable predictive gains while penalizing fragility and cost.

Large-scale systems also introduce infrastructure constraints. Distributed joins and group-by operations may dominate runtime; repeated materialization of intermediate features increases storage expense; and inconsistent execution across training and serving environments creates training–serving skew. Efficient frameworks need lazy evaluation, caching, partition-aware aggregation, deterministic transformations, and reusable metadata. For real-time applications, feature freshness and inference latency may matter as much as batch model accuracy. A pipeline must therefore be evaluated as an operational system, not merely as a model fitted once.

Predictive reliability extends beyond average accuracy. Models should remain calibrated, robust under realistic missingness and drift, stable across repeated splits, and interpretable enough for review. Evaluation should report task-appropriate metrics, uncertainty intervals, and performance for important subgroups. Temporal datasets require chronological validation; entity-linked observations require group-aware separation. Without these controls, feature searches can overfit the validation set and create inflated estimates of deployment performance.

This paper proposes a unified architecture that joins automated profiling, policy-driven repair, feature synthesis, constrained feature selection, model training, and continuous monitoring. Its primary contribution is a design specification with explicit quality metrics, optimization equations, and an auditable workflow. It further outlines an evaluation protocol that can establish whether integration improves predictive quality, stability, and efficiency over conventional baselines. The numerical example is deliberately labeled illustrative so that the distinction between a proposed method and experimentally established evidence remains clear.

Related Work

Foundational data-quality research established that quality is dependent on use and cannot be captured adequately by syntactic correctness alone. Wang and Strong organized dimensions of data quality from the perspective of information consumers, motivating approaches that incorporate accessibility, contextual relevance, and representational clarity. Pipino, Lee, and Wang subsequently argued for practical assessment procedures that combine subjective and objective measures. These contributions remain relevant because predictive systems need a task-specific interpretation of what makes an input trustworthy.

Database research on data cleaning examined integrity constraints, entity resolution, duplicate detection, and error correction. Rule-based approaches provide transparency and are particularly useful for schema violations, impossible ranges, and conflicting records. Their principal limitation is coverage: unanticipated errors or complex dependencies may evade predefined checks. Statistical profiling and learned constraints extend coverage, but they can misclassify legitimate rare cases. Modern pipelines therefore benefit from combining deterministic rules with probabilistic anomaly scores and review mechanisms.

The literature on missing data distinguishes mechanisms commonly described as missing completely at random, missing at random, and missing not at random.

Little and Rubin emphasize that the validity of an imputation strategy depends on assumptions about these mechanisms. In applied machine learning, simple imputation remains widespread because it is fast and reproducible, while iterative and model-based methods may preserve relationships more effectively under appropriate conditions. No method is universally superior; selection should be guided by controlled masking experiments and task performance.

Robust statistics and anomaly detection provide complementary strategies for identifying unusual observations. Median absolute deviation and quantile-based rules resist the influence of extreme values. Isolation Forest, introduced by Liu, Ting, and Zhou, identifies anomalies using unusually short random partition paths. Such techniques are useful for scalable profiling, but anomaly scores are not proof of incorrectness. Research on imbalanced classification also warns that rare observations may contain disproportionate predictive information, particularly when positive outcomes are scarce.

Feature selection research distinguishes filter methods, wrappers, and embedded techniques. Filters rank candidates using relevance measures; wrappers estimate model performance for feature subsets; embedded approaches select or shrink features during training. Guyon and Elisseeff describe the importance of evaluating relevance, redundancy, and selection stability. High-dimensional settings make this balance difficult because seemingly predictive features can arise by chance. Nested validation and stability checks reduce the risk of optimistic estimates from repeated selection.

Automated feature engineering systems have explored systematic transformation search. Deep Feature Synthesis, described by Kanter and Veeramachaneni, composes aggregation and transformation primitives across relational data. Other AutoML research has addressed end-to-end model and hyperparameter search. These approaches reduce manual effort but require careful search-space design and compute budgeting. For deployment, feature discovery must additionally consider transformation reproducibility, freshness, and availability at prediction time.

Tree-based learning methods such as random forests and gradient boosting often handle nonlinear interactions without explicit cross-product expansion. Breiman demonstrated the effectiveness of ensemble trees, while Chen and Guestrin developed XGBoost for scalable boosted-tree training. Even powerful learners, however, can suffer from faulty timestamps, unstable categories, leakage, or systematic missingness. This suggests that data-quality governance remains valuable irrespective of model family.

Data drift and dataset shift research investigates changes between training and deployment distributions. Covariate shift, prior-probability shift, and concept shift have different implications for retraining and monitoring. Gama and colleagues survey concept-drift adaptation, including detection and response strategies. Distributional tests can provide early warnings, but detecting a changed feature does not establish a corresponding decline in predictive performance. Effective monitoring links drift alarms with delayed labels, model calibration, and operational outcomes.

Interpretability and fairness studies raise further concerns about automated transformations. A derived variable can encode sensitive proxies or become difficult to explain, while aggregate performance may conceal substantial subgroup variation. Local explanation techniques, such as SHAP, can help examine model contributions, although explanations do not by themselves establish causality or data correctness. Feature provenance, transformation documentation, and subgroup audits should therefore accompany predictive metrics.

Across these research streams, a recurring gap is the separation of quality assessment from feature-search objectives. Data-cleaning methods often optimize conformity or reconstruction, while feature engineering methods optimize predictive score. An integrated system can test whether a proposed correction genuinely improves out-of-sample utility and whether a generated feature remains reliable under plausible data perturbations. The present framework builds on established components rather than claiming a new standalone learning algorithm, and its effectiveness must be validated through transparent comparisons on real datasets.

Proposed Methodology

The proposed Quality-Aware Automated Feature Engineering (QAAFE) architecture consists of ingestion, profiling, risk-aware remediation, candidate feature generation, constrained selection, predictive validation, and monitoring. Data partitions are defined before any learned preprocessing. Each transformation is versioned with its training statistics and feature lineage.

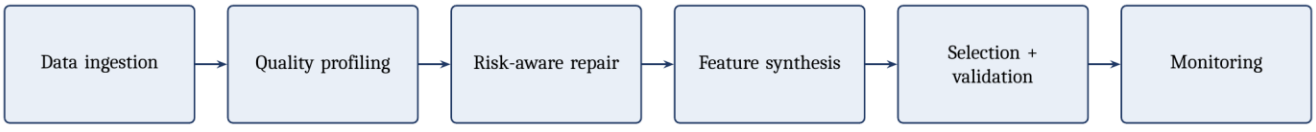


Figure 1: Proposed quality-aware automated predictive analytics workflow.

Data-quality scoring

Let X be a data matrix with n observations and d variables. For each variable j, define completeness Cj = 1 − mj/n, where mj is the count of missing values. Let Vj be the proportion satisfying domain constraints, Uj the proportion satisfying applicable uniqueness checks, and Sj a normalized stability score computed against a training reference. An aggregate score is:

$$Qj = wC\,Cj + wV\,Vj + wU\,Uj + wS\,Sj, \text{ with } wk \geq 0 \text{ and } \Sigma wk = 1 \quad (1)$$

Weights reflect business risk and must be documented. Since some dimensions are inapplicable to particular variables, weights are renormalized over applicable checks. A feature-level quality score is not treated as an estimate of ground-truth accuracy.

Risk-aware remediation

Candidate repair actions include leaving values unchanged with missingness flags, robust median or mode imputation, conditional imputation, category normalization, duplicate consolidation, and targeted quarantine. The action is selected on training folds by minimizing expected predictive loss plus penalties for repair cost and distributional distortion:

$$a^* = \text{argmin}_a\,[\,Lval(fa) + \lambda 1\,Cost(a) + \lambda 2\,Distortion(a)\,] \quad (2)$$

Corrections with high potential impact are logged for human review. Invalid or extreme observations are never automatically deleted solely because an anomaly detector flags them.

Intelligent feature generation and selection

A constrained grammar creates candidate transformations φk(X), including time-safe lags, counts, ratios with guarded denominators, logarithmic transforms, and interactions. Candidates must be computable at inference time and preserve entity and temporal boundaries. The selected feature subset S solves:

$$S^* = \text{argmax}_S\,[\,AUROCcv(S) - \alpha|S|/K - \beta Latency(S) - \gamma Instability(S) - \delta Risk(S)\,] \quad (3)$$

Here K is the number of candidates, AUROCcv is cross-validated discrimination, and Instability measures variation in feature relevance across folds. For regression tasks, a suitably oriented loss or R^2 can replace AUROC. Selection takes place inside each training fold to avoid leakage.

Model validation and operational monitoring

Train, validation, and holdout data are separated by time or entity where appropriate. The primary baseline uses a conventional imputation-and-encoding pipeline; additional baselines isolate quality repair and feature engineering. The final model is evaluated once on an untouched holdout. Confidence intervals are estimated using entity-aware bootstrap sampling. Drift monitoring may use population stability index (PSI):

$$PSI = \Sigma b\,(pb - qb)\,\ln(pb/qb) \quad (4)$$

where pb and qb are reference and current proportions in bin b, after smoothing zero counts. PSI is a descriptive drift indicator rather than a calibrated hypothesis test. Retraining requires corroborating evidence of performance deterioration or material changes in data-generating processes.

Results and Discussion — Illustrative Scenario

The following values are a constructed demonstration of how an empirical comparison could be presented. They are not derived from a collected dataset, executed experiments, or validated measurements. The scenario assumes a binary classification problem with the same held-out population and comparable computing resources for each configuration.

Table 1: Hypothetical comparative performance of pipeline configurations.

Pipeline	*Accuracy (%)*	*F1 (%)*	*AUROC*	*Runtime (min)*
Basic preprocessing	82.4	79.8	0.871	24
Quality assessment + repair	85.1	82.6	0.896	31
Feature engineering only	86.3	84.0	0.908	39
Integrated QAAFE	89.2	87.1	0.934	47

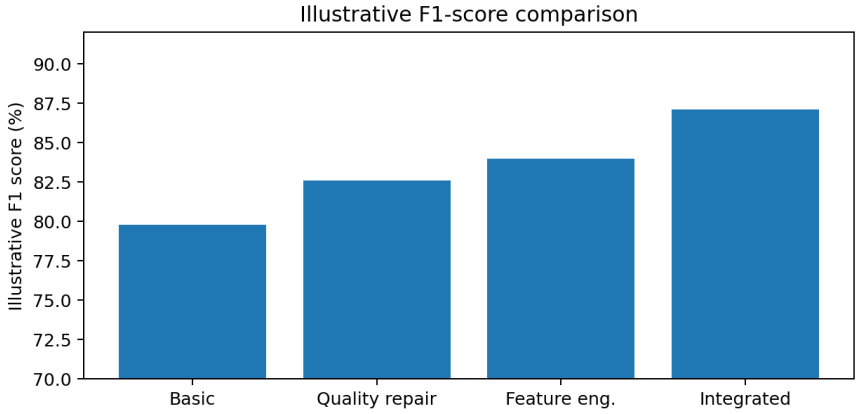


Figure 2: Illustrative F1-score values; not empirical measurements.

In this hypothetical scenario, integrated processing improves F1 by 7.3 percentage points over basic preprocessing and AUROC by 0.063, while increasing processing time by 23 minutes. This pattern demonstrates a plausible accuracy–cost tradeoff rather than evidence of actual superiority. The additional compute burden motivates profiling-based early rejection of low-value features, incremental caching, and explicit latency budgets.

An empirical study should repeat the comparison on multiple publicly documented datasets with different missingness patterns, cardinalities, and drift profiles. It should report confidence intervals, training and inference latency separately, peak memory, failure rates, and per-group performance. Ablations should remove the quality gate, repair policy, stability penalty, and temporal-leakage constraints one at a time. Statistical claims require actual repeated measurements and appropriate tests.

Conclusion

This paper specifies an integrated framework for automated data-quality assessment and intelligent feature engineering in large-scale predictive analytics. The architecture combines multidimensional profiling, risk-aware remediation, provenance-preserving feature synthesis, constrained selection, leakage-safe evaluation, and ongoing monitoring. Its central proposition is that predictive utility and data trustworthiness should be optimized jointly rather than handled as isolated stages. The illustrative comparison shows how improvements and computational tradeoffs could be reported but does not establish empirical effectiveness. Future work should implement the system on distributed infrastructure, benchmark it across diverse real datasets, assess calibration and subgroup robustness, and evaluate whether automated quality interventions consistently improve deployment reliability.

References

1. Rangappa, A. S. (2020). Integration of behavioral analytics in developer workflows for detecting insider threats and malicious activities through activity monitoring and contextual profiling. International Journal of Innovative Research in Science, Engineering and Technology, 9(1), 13375–13383. https://doi.org/10.15680/IJIRSET.2020.0901064
2. Talasila, D. (2018). Utilization of runtime application self-protection (RASP) in production environments for real-time vulnerability mitigation and behavioral enforcement through embedded monitoring agents. International Journal of Advanced Research in Education and Technology, 5(3), 831–840. https://doi.org/10.15680/IJARETY.2018.0503017
3. Aggarwal, A. (2018). Implementation of cloud security automation in continuous integration pipelines for accelerating DevSecOps adoption and minimizing manual misconfigurations through security-as-code. International Journal of Innovative Research in Science, Engineering and Technology, 7(11). https://doi.org/10.15680/IJIRSET.2018.0711083
4. Mahadasu, S. S. (2025). Intent-driven MVNO architecture for cross-border 5G/6G networks in Africa: Toward the One Africa digital economy. In 2025 IEEE Connecting the Unconnected Europe, Middle East, and Africa Regional Summit (CTUS-EMEA). IEEE. https://doi.org/10.1109/CTUS-EMEA68989.2025.11592181
5. Kumar, Ashok, and Arpit Jain. "Image smog restoration using oblique gradient profile prior and energy minimization." Frontiers of Computer Science 15, no. 6 (2021): 156706.
6. Anchala, S. (2018). Adoption of infrastructure as code security validation in cloud-native environments for reducing configuration drifts and compliance violations through automated template scanning. International Journal of Innovative Research in Science, Engineering and Technology, 7(10), 8305–8313. https://doi.org/10.15680/IJIRSET.2018.0710087
7. Talasila, D. (2018). Implementation of versioned configuration management in DevSecOps for preventing misconfigurations and ensuring environmental consistency through GitOps automation. International Journal of Innovative Research in Computer and Communication Engineering, 6(2), 1731–1739. https://doi.org/10.15680/IJIRCCE.2018.0602134
8. Khan, S. (2025). The role of artificial intelligence in cyber security: Leveraging machine learning and predictive analytics for intrusion detection, threat intelligence, and autonomous defense mechanisms. The Research Journal, 11(3), 1–7.
9. Talasila, D. (2019). DevSecOps in e-governance platforms: Safeguarding public data, ensuring transparency, and enhancing trust in cloud-based government applications. International Journal of Innovative Research in Science, Engineering and Technology, 8(2), 1561–1570. https://doi.org/10.15680/IJIRSET.2019.0802107
10. Ahuja, R. (2022). Integration of runtime application self-protection (RASP) in web applications for real-time threat mitigation and code behavior monitoring through embedded security instrumentation.

NeuroQuantology, 20(11), 10202–10217. https://doi.org/10.48047/nq.2022.20.11.nq67013

11. Anchala, S. (2018). Deployment of container image scanning in DevSecOps pipelines for mitigating open-source vulnerabilities and ensuring runtime safety through CVE-based detection mechanisms. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 7(3), 1546–1554. https://doi.org/10.15662/IJAREEIE.2018.0703058
12. Aggarwal, A. (2019). Application of secure Virtual Private Cloud (VPC) architectures in public cloud infrastructure for strengthening network isolation and segmenting trust zones through fine-tuned routing. International Journal of Multidisciplinary and Scientific Emerging Research, 7(1), 423–431. https://doi.org/10.15662/IJMSERH.2019.0701048
13. Talasila, D. (2018). Integration of identity federation in multi-cloud access management for streamlining user authentication and reducing credential sprawl through centralized SSO protocols. International Journal of Multidisciplinary and Scientific Emerging Research, 6(4), 2017–2025. https://doi.org/10.15662/IJMSERH.2018.0604046
14. Rao, S. Madhusudhana, and Arpit Jain. "Advances in Malware Analysis and Detection in Cloud Computing Environments: A Review." International Journal of Safety & Security Engineering 14, no. 1 (2024).
15. Aggarwal, A. (2018). Application of federated learning in medical IoT devices for preserving patient privacy and preventing data breaches through decentralized cyber defense models. International Journal of Innovative Research in Computer and Communication Engineering, 6(3), 2654–2660. https://doi.org/10.15680/IJIRCCE.2018.0603167
16. Bhanushali, B. (2022). Blockchain-driven carbon credit trading systems for improving environmental transparency, reducing double counting, and ensuring accountability in climate finance. International Journal on Recent and Innovation Trends in Computing and Communication, 10(11), 292–300. https://www.ijritcc.org/index.php/ijritcc/article/view/11927
17. Bhaskara Raju Rallabandi. (2022). DETERMINISTIC STABILITY ANALYSIS FOR PRODUCTION-LINE CONTROL OVER PRIVATE 5G. American Journal of AI Cyber Computing Management, 2(4), 33–37. https://doi.org/10.64751/ajaccm.2022.v2.n4.pp33-37
18. Jain, Arpit, Nageswara Rao Moparthi, A. Swathi, Yogesh Kumar Sharma, Nitin Mittal, Ahmed Alhussen, Zamil S. Alzamil, and MohdAnul Haq. "Deep Learning-Based Mask Identification System Using ResNet Transfer Learning Architecture." Computer Systems Science & Engineering 48, no. 2 (2024): 341.
19. Bhaskara Raju Rallabandi. (2023). NATIVE OBSERVABILITY FOR PRIVATE 5G: A REFERENCE ARCHITECTURE ON COMBINATION OF NWDAF, SMO, AND EDGE ANALYTICS. American Journal of AI Cyber Computing Management, 3(1), 24–28. https://doi.org/10.64751/ajaccm.2023.v3.n1.pp24-28
20. Bhanushali, B. (2023). Blockchain for trade finance: Enhancing cross-border payment transparency, reducing processing time, and mitigating fraud through distributed ledger systems. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 5690–5697. https://www.ijritcc.org/index.php/ijritcc/article/view/11932
21. Dwivedi, D. (2023). Integration of secure file upload mechanisms in user-facing applications for preventing malware injections and file-based exploits through content-type validation. International Journal on Recent and Innovation Trends in Computing and Communication, 11(6), 768–775. https://www.ijritcc.org/index.php/ijritcc/article/view/11940
22. Anchala, S. (2019). Database security challenges in cross-border data flows: Analysing compliance, privacy, and risk in international data transfers. International Journal of Advanced Research in Education and Technology, 6(3), 831–838. https://doi.org/10.15680/IJARETY.2019.0603016
23. Rangappa, A. S. (2025). Agentic AI and cyber security: Autonomous threat hunting, intrusion detection, and adaptive defense mechanisms in a world of increasingly sophisticated cyber attacks. Journal of Digital Security and Forensics, 2(1), 128–138. https://doi.org/10.29121/digisecforensics.v2.i1.2025.86
24. Jain, Arpit, Adesh Kumar, and Sanjeev Sharma. "Comparative design and analysis of mesh, torus and ring NoC." Procedia Computer Science 48 (2015): 330–337.
25. Dwivedi, D. (2020). Adoption of secure software updates in application lifecycle for preventing supply chain tampering and maintaining trust through cryptographically signed patches. Journal of Critical Reviews, 7(10), 6974–6980. https://doi.org/10.69980/jcr.v7:i10.13406
26. Anchala, S. (2019). Database security in e-governance systems: Protecting citizen records, ensuring transparency, and managing public trust in cloud-based platforms. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 2(11), 2268–2276. https://doi.org/10.15680/IJMRSET.2019.0211014
27. Kumar, S., Sharma, K., Kumar, P. A., Jain, A., Bhagat, S. K., & Singh, P. (2024, September). An improved particle swarm approach for energy-aware location-aided routing in mobile ad-hoc network. In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 7, pp. 1119–1124). IEEE.
28. Gupta, A. (2025). Agentic AI for national security and defense: An analytical study on autonomous agents in surveillance, strategic decision-making, and military operations with a focus on ethical and legal boundaries. Journal of Digital Security and Forensics, 2(1), 163–175. https://doi.org/10.29121/digisecforensics.v2.i1.2025.84
29. Mahadasu, S. S. (2025). Distributed AI-RAN inference placement for joint radio optimization and edge workload acceleration. International Journal of AI Electronics and Nexus Energy, 1(1), 39–44. https://doi.org/10.64751/7ep35m50
30. Devulapalli, N. (2024). Agentic AI in finance and economics: The role of autonomous agents in fraud detection, algorithmic trading, credit risk assessment, and ethical governance of automated financial systems. International Journal of Multidisciplinary and Scientific Emerging Research, 12(4), 1958–1964. https://doi.org/10.15662/IJMSERH.2024.1204057
31. Ahuja, R. (2024). Deployment of AI in robotic process automation: Enhancing enterprise efficiency and minimizing manual errors via intelligent workflow automation. International Journal of Advanced Research in Education and Technology, 11(3), 1330–1337. https://doi.org/10.15680/IJARETY.2024.1103068
32. Devulapalli, N. (2025). Machine learning in database security: Leveraging anomaly detection, intrusion detection, and predictive analytics for proactive threat prevention. International Journal of Innovative Research in Science, Engineering and Technology, 14(2), 1685–1692. https://doi.org/10.15680/IJIRSET.2025.1402105
33. Gupta, A. (2021). Application of cyber deception strategies in military networks for enhancing operational security and misleading attackers through adaptive honeypots and decoy systems. NeuroQuantology, 19(4), 317–333. https://doi.org/10.48047/nq.2021.19.4.NQ21068
34. Devulapalli, N. (2024). The role of homomorphic encryption in database security: Enabling confidential query processing, secure data sharing, and privacy-preserving analytics. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(8), 14319–14326. https://doi.org/10.15680/IJMRSET.2024.0708082
35. Chatrath, A. (2018). The impact of containerization on DevSecOps: Security considerations for Docker, Kubernetes, and microservices architectures in continuous deployment pipelines. International Journal of Advanced Research in Education and Technology, 5(6). https://doi.org/10.15680/IJARETY.2018.0506019
36. Gupta, Namit, Kunwar Singh Vaisla, Arpit Jain, Adesh Kumar, and Rajeev Kumar. "Performance Analysis of AODV Routing for Wireless Sensor Network in FPGA Hardware." Computer Systems Science & Engineering 40, no. 3 (2022).
37. Chatrath, A. (2019). A multi-layered DevSecOps framework: Integrating secure coding, automated testing, vulnerability scanning, and incident response into end-to-end software development lifecycles. International Journal of Advanced Research in Electrical,

Electronics and Instrumentation Engineering, 8(3). https://doi.org/10.15662/IJAREEIE.2019.0803102

38. Mahadasu, S. S. (2025). Semantic observability for autonomous telecom operations using graph reasoning and multi-agent decision validation. International Journal of AI Electronics and Nexus Energy, 1(3), 25–30. https://doi.org/10.64751/6ppk3q13
39. Gupta, A. (2020). Utilization of homomorphic encryption in healthcare data sharing for ensuring confidentiality and compliance through secure computation without data exposure. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 3(10). https://doi.org/10.15680/IJMRSET.2020.0310011
40. Ahuja, R. (2019). Integrating security testing into CI/CD pipelines: An analytical study on static application security testing (SAST), dynamic application security testing (DAST), and interactive application security testing (IAST) in DevSecOps. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(2), 491–499. https://doi.org/10.15662/IJAREEIE.2019.0802045
41. Bhaskara Raju Rallabandi. (2024). DIGITAL TWIN–DRIVEN 6G CAMPUS NETWORKS: INTELLIGENT FAULT RECOVERY AND ENERGY OPTIMIZATION. American Journal of AI Cyber Computing Management, 4(3), 30–34. https://doi.org/10.64751/ajaccm.2024.v4.n3.pp30-34
42. Harshitha, G., Kumar, S., Rani, S., & Jain, A. (2021, November). Cotton disease detection based on deep learning techniques. In 4th Smart Cities Symposium (SCS 2021) (Vol. 2021, pp. 496–501). IET.
43. Mahadasu, S. S. (2025). Distributed AI and ISAC for autonomous agricultural infrastructure: Toward cognitive 6G farming ecosystems. International Journal of Engineering Research and Science & Technology, 21(2), 3110–3116. https://doi.org/10.62643/ijerst.2025.v21.n2.4403
44. Mahajan, R. A., Dey, R., Khan, M., Mohd Su'ud, M., Alam, M. M., & Jadhav, P. (2025). Enhancing personalization in IoT-based health monitoring via generative AI and transfer learning. Egyptian Informatics Journal, 32, Article 100788. https://doi.org/10.1016/j.eij.2025.100788
45. Bhanushali, B. (2020). The industry needs a validated model to justify the immense cost of migrating from manual, periodic KYC (know your customer) to automated. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 3(11), 1799–1805. https://doi.org/10.15680/IJMRSET.2020.0311009
46. Jain, A. (2024, December). Energy-Efficient Signal Processing Architectures for IoT Networks Design. In 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN) (pp. 362–367). IEEE.
47. Aggarwal, A. (2018). Implementation of quantum cryptography in financial transactions for enhancing data confidentiality and resisting future quantum attacks through secure key distribution mechanisms. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 7(7), 3196–3204. https://doi.org/10.15662/IJAREEIE.2018.0707016
48. Bhaskara Raju Rallabandi. (2022). SUB-MICROSECOND 5G SYNC FOR RELIABLE FACTORY ROBOT CONTROL. American Journal of Management and IOT Medical Computing, 1(1), 34–39. https://doi.org/10.64751/ajmimc.2022.v1.n1.pp34-39
49. Kumar, S., Jain, A., & Moparthi, N. R. (2022). Enhanced method of object tracing using extended Kalman filter via binary search algorithm.
50. Mahajan, R. A., Dey, R., Khan, M., Kanabar, K., & Gupta, S. K. (2025). Real-time vehicular-to-vehicular (V2V) communication and applications for pedestrian safety in AI-optimized VANETs for autonomous vehicles. In H. Maryam, M. M. Malik, I. U. Khan, & S. K. Gupta (Eds.), AI-driven transportation systems: Real-time applications and related technologies (pp. 151–167). Springer. https://doi.org/10.1007/978-3-031-98349-8_8
51. Chatrath, A. (2020). DevSecOps for financial services: An in-depth study of secure CI/CD practices, regulatory compliance, and real-time threat monitoring in cloud-based FinTech applications. International Journal of Innovative Research in Science, Engineering and Technology, 9(11). https://doi.org/10.15680/IJIRSET.2020.0911098
52. Khan, S. (2023). AI-powered innovations in cross-border merchant payment systems: Exploring the role of artificial intelligence in enhancing speed, transparency, and trust in international financial transactions. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 12(5), 1217–1224. https://doi.org/10.15662/IJAREEIE.2023.1205026
53. Ahuja, R. (2021). A study of multi-layered database security architectures: Combining access control, encryption, and intrusion prevention for comprehensive protection. International Journal on Recent and Innovation Trends in Computing and Communication, 9(2), 42–49. https://www.ijritcc.org/index.php/ijritcc/article/view/11920
54. Rangappa, A. S. (2018). Deployment of immutable infrastructure in Dev Sec Ops practices for minimizing configuration drift and recovery time through ephemeral build environments. International Journal of Multidisciplinary and Scientific Emerging Research, 6(4), 2026–2034. https://doi.org/10.15662/IJMSERH.2018.0604047
55. Khan, S. (2023). AI-driven speech and voice analytics in CRM platforms: A study on intelligent call center automation, emotion recognition, and service personalization. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 6(2), 510–517. https://doi.org/10.15680/IJMRSET.2023.0602035
56. Jain, Jaishree, Arpit Jain, Saurabh Kumar Srivastava, Chaman Verma, Maria Simona Raboaca, and Zoltán Illés. "Improved security of e-healthcare images using hybridized robust zero-watermarking and hyper-chaotic system along with RSA." Mathematics 10, no. 7 (2022): 1071.
57. Dwivedi, D. (2022). Application of AI-powered vulnerability scanners in legacy applications for identifying unpatched security flaws and weak dependencies through semantic code analysis. International Journal on Recent and Innovation Trends in Computing and Communication, 10(12), 617–623. https://www.ijritcc.org/index.php/ijritcc/article/view/11918

HOW TO CITE THIS ARTICLE: Tursunov, A., & Khan, M. A. (2025). Automated Data-Quality Assessment and Intelligent Feature Engineering for Large-Scale Predictive Analytics. *Journal of Integrated Science, Technology and Management*, 1(2), 15-21.