

Integrated Feasibility Assessment of Edge, Cloud, and Hybrid AI-as-a-Service Architectures for Face Recognition

Pradipa Catya Wisesa^{*1}, Ary Murti^{*2}, Helni M.J^{*3}

^{1,2,3} School of Electrical Engineering, Telkom University, Bandung 40257, Indonesia

Article Info

Article history:

Received July 30, 2026

Revised Sep 07, 2026

Accepted Oct 05, 2026

Keywords:

AI-as-a-Service

Edge computing

Face recognition

Feasibility study

Hybrid architecture

ABSTRACT

Deploying face recognition as a service requires evidence beyond recognition accuracy because architecture affects latency, scalability, cost, and biometric-data exposure. This study evaluates Edge, Cloud, and Hybrid architectures through an integrated technical, techno-economic, and regulatory-readiness framework. EdgeCloudSim experiments modeled 100–500 connected devices using an infrastructure-level workload abstraction structurally mapped to published face-recognition pipelines, while explicitly distinguishing this abstraction from direct neural-inference benchmarking. The technical screening thresholds were treated as scenario-specific engineering gates and evaluated through threshold sensitivity analysis. A five-year comparative cash-flow model applied the same modeled avoided operating-cost benefit to all architectures. At 500 devices, Hybrid produced 2.2066 s processing time, 16.736% failed tasks, and 0.1355 s total latency; its failure rate was lower than Edge (27.797%) and Cloud (22.137%). Under the common economic-benefit scenario, Edge and Cloud generated negative NPVs of IDR 982.50 million and IDR 247.25 million, respectively, whereas Hybrid generated an NPV of IDR 958.96 million, IRR of 30.73%, PI of 1.64, and an actual cumulative payback period of 2.60 years. Regulatory mapping assessed implementation readiness rather than demonstrated legal compliance. Hybrid therefore remains the most balanced conditional baseline under the stated assumptions, subject to empirical pipeline calibration, capacity upgrades, privacy controls, and field validation.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Pradipa Catya Wisesa

School of Electrical Engineering, Telkom University Bandung 40257, Indonesia

Email: pradipawisesa@student.telkomuniversity.ac.id, arymurti@telkomuniversity.ac.id, helnimj@telkomuniversity.ac.id

1. INTRODUCTION

Face recognition has moved from laboratory demonstrations to operational services for access control, attendance, border processing, and security monitoring. Recent reviews show that convolutional neural networks and embedding-based matching have substantially improved recognition capability, while deployment studies increasingly emphasize real-time performance, environmental variation, and implementation context (EL Fadel, 2025; Hidayat et al., 2024). Adoption nevertheless remains sociotechnical: convenience and automation must be weighed against privacy, perceived risk, and automation bias (Bhatia & Mathur, 2026). Consequently, a high benchmark accuracy does not by itself establish that a face recognition system is feasible for enterprise use.

The location of inference is a central design decision. Cloud deployment offers elastic computation and centralized management, but transmitting image or video workloads through a wide-area network can create bandwidth dependence and end-to-end delay. Edge deployment places computation near the camera and can reduce network latency and raw-data movement, although embedded resources constrain memory, processing capacity, and energy efficiency (Merenda et al., 2020; Satyanarayanan, 2017; T. Wang et al., 2025). Hybrid deployment divides the pipeline: local nodes perform capture, frame selection, or face extraction, whereas cloud resources perform more intensive identification and centralized storage. The resulting design problem is therefore a trade-off among compute saturation, network transmission, service reliability, and privacy exposure.

Prior technical work provides strong evidence for this trade-off. Koubaa et al. (2022) compared cloud, edge, and hybrid face-recognition inference and documented architecture-dependent differences

in execution time, energy, and memory. Oroceo et al. (2022) showed that collaborative edge–cloud inference can outperform isolated deployment strategies. Broader video-analytics research similarly demonstrates that preprocessing, filtering, compression, and adaptive offloading can reduce network load, although aggressive data reduction may introduce accuracy trade-offs (Badidi et al., 2023; Nguyen & Choi, 2020; Y. Wang et al., 2023; Xu et al., 2023). These studies explain technical performance, but they do not determine whether a selected architecture is simultaneously affordable and lawful in the Indonesian biometric-data context.

Feasibility research offers a complementary perspective. Ssegawa & Muzinda (2021) proposed a systematic project assessment framework, while Jacob et al. (2025) argued that real-world AI assessment should incorporate integration, governance, cost, safety, scalability, and impact. Corporate AI research likewise reports that implementation value depends on organizational and economic conditions, not only algorithmic output (Díaz Martínez et al., 2024). However, general feasibility frameworks do not operationalize the data flows and performance bottlenecks of face recognition, and domain-specific AI frameworks are not direct substitutes for a deployment comparison under Indonesian law.

A second limitation concerns responsible operation. Demographic performance differences and bias have been documented across facial-analysis systems, making field validation and human oversight necessary even when the infrastructure itself performs adequately (Cavazos et al., 2021; Khalil et al., 2020). Biometric templates also remain sensitive data even when raw video is not retained. Architecture can reduce exposure, but cannot by itself supply a lawful basis, transparency, retention governance, security, or accountability. These considerations reinforce the need for an integrated decision framework rather than a ranking based on a single speed or cost indicator.

This study addressed that gap by comparing Edge, Cloud, and Hybrid AI-as-a-Service architectures using one set of workload assumptions and three linked dimensions: technical performance, techno-economic viability, and regulatory readiness. Its novelty lies in tracing the deployment decision from simulated infrastructure behavior to cost structure and the obligations created by biometric-data flows. The face-recognition specificity of the simulation is represented through the placement of detection, face extraction, embedding/identification, and network-transfer stages reported in prior deployment studies; the simulation is not presented as a direct benchmark of a particular neural network or camera. The study asked which architecture provided the most defensible overall baseline for enterprise attendance and security applications. It hypothesized that Hybrid would offer the strongest balance by limiting raw-video transmission without concentrating the entire inference workload on constrained local resources. Table 1 positions this contribution against representative prior studies.

Table 1. State of the Art and Research Position

Study	Primary focus	Limitation relative to this study
Koubaa et al. (2022)	Cloud, Edge, and Hybrid inference performance	No integrated investment and Indonesian regulatory assessment
Oroceo et al. (2022)	Collaborative edge–cloud face inference	Technical optimization without lifecycle feasibility integration
Ssegawa & Muzinda (2021)	General project feasibility framework	Not operationalized for AI inference or biometric data flows
Jacob et al. (2025)	Multidimensional real-world AI evaluation	Healthcare-specific; no architecture simulation
Present study	Technical simulation + finance + PDP-law mapping	Integrated, deployment-oriented comparison for enterprise face recognition

2. METHOD

This research used a comparative, simulation-based feasibility design. The unit of analysis was the deployment architecture rather than an individual user. Three scenarios were evaluated: (1) Edge, in which face detection and embedding/identification were executed near the camera and only results or metadata were forwarded; (2) Cloud, in which the captured frame workload was transmitted through the WAN and detection plus recognition were processed centrally; and (3) Hybrid, in which local nodes performed frame selection, face detection, and cropping before a reduced facial payload was sent to the cloud for identification. No human participants or live biometric records were used. Accordingly, the technical experiment evaluates infrastructure-level resource and network behavior of a face-recognition service abstraction; it does not measure recognition accuracy, demographic performance, or the exact per-frame inference time of a specific detector or recognizer.

Technical experiments were configured in EdgeCloudSim, an environment designed to evaluate resource and network behavior in edge-computing systems (Sonmez et al., 2018). Device load increased from 100 to 500 in increments of 100. A NEXT_FIT orchestration policy was retained across single-tier and two-tier scenarios so that observed differences could be attributed primarily to placement of computation and network traversal. The centralized Cloud consisted of one host with four virtual machines, each allocated four cores, 10,000 MIPS, and 52,000 MB RAM. The distributed Edge environment comprised 14 data centers, with one host per data center and two virtual machines per host; each virtual machine had two cores, 1,000 MIPS, and 2,000 MB RAM. The workload was calibrated structurally, rather than numerically, against published face-recognition deployment pipelines. Koubaa et al. (2022) evaluated MTCNN-based face detection and FaceNet recognition under cloud, edge, and hybrid deployment, while Oroceo et al. (2022) implemented local MTCNN face detection with cropped-face transfer to cloud-side recognition. These studies provide empirical support for the stage-placement and payload-flow logic used here, but their hardware-specific inference times were not imposed on EdgeCloudSim. Table 2 therefore distinguishes architectural mapping from direct model-level calibration. from 100 to 500 in increments of 100. A NEXT_FIT orchestration policy was retained across single-tier and two-tier scenarios so that observed differences could be attributed primarily to placement of computation and network traversal. The centralized Cloud consisted of one host with four virtual machines, each allocated four cores, 10,000 MIPS, and 52,000 MB RAM. The distributed Edge environment comprised 14 data centers, with one host per data center and two virtual machines per host; each virtual machine had two cores, 1,000 MIPS, and 2,000 MB RAM. Table 2 summarizes the parameters used for comparison.

Table 2. Core EdgeCloudSim Configuration and Comparison Logic

Parameter	Configuration
Workload	100, 200, 300, 400, and 500 connected devices
Orchestration	NEXT_FIT; SINGLE_TIER and TWO_TIER scenarios
Cloud compute	1 host; 4 VMs; 4 cores/VM; 10,000 MIPS and 52,000 MB RAM/VM
Edge compute	14 data centers; 2 VMs/host; 2 cores/VM; 1,000 MIPS and 2,000 MB RAM/VM
Network	WLAN bandwidth 200; WAN bandwidth 15; WAN delay 0.1; LAN delay 0.005
Metrics	Processing time, failed-task percentage, and total latency
Decision targets	Processing ≤ 2 s; total latency ≤ 3 s; failed tasks $\leq 0.1\%$ (scenario-specific screening gates)
Face-recognition workload mapping	Edge: detection + embedding/identification local; Cloud: full-frame WAN transfer + central detection/recognition; Hybrid: local detection/cropping + reduced facial payload to cloud identification. Stage placement follows Koubaa et al. (2022) and Oroceo et al. (2022).
Calibration scope	Structural pipeline calibration only; EdgeCloudSim processing time represents simulated service-task time and is not claimed to equal hardware-specific MTCNN/FaceNet/Dlib inference latency. Recognition accuracy is outside the simulation scope.

Three outcome metrics were recorded for every device load. Processing time represented compute-side service time; failed-task percentage represented requests not completed because of resource or queue constraints; and total latency captured end-to-end delay. The study retained screening gates of ≤ 2 s processing time, ≤ 3 s total latency, and $\leq 0.1\%$ failed tasks. These values are explicitly treated as scenario-specific engineering gates rather than universal standards, regulatory requirements, or an enterprise SLA. The 2 s processing and 3 s end-to-end limits define a conservative interactive-service envelope consistent with the real-time emphasis of prior edge/cloud face-recognition deployment research (Koubaa et al., 2022; Oroceo et al., 2022) and with the broader importance of low network latency in interactive systems (Parvez et al., 2018). The 0.1% failed-task gate represents a deliberately stringent service-reliability objective. Because no organization-specific SLA was available and the literature does not establish a universal triplet of thresholds for this use case, the results were additionally tested at the maximum workload against alternative processing limits of 1.5, 2.0, 2.5, and 3.0 s; latency limits of 0.2, 0.5, 1.0, and 3.0 s; and failed-task limits of 0.1%, 1%, 5%, 10%, 20%, 25%, and 30%. This sensitivity analysis was used to determine whether the architecture interpretation depended on a single arbitrary threshold.

The techno-economic stage compared architecture-specific capital expenditure (CapEx) and first-year operating expenditure (OpEx) for a baseline enterprise deployment of 14 locations and 100 cameras. To address comparability across architectures, the same economic-benefit model was applied to Edge, Cloud, and Hybrid. The modeled Year-1 benefit of IDR 850 million is defined as an avoided operating-cost scenario associated with replacing manual identity verification and related attendance/security administration for the same service scope; it is not subscription revenue and is not presented as an observed accounting saving from a specific organization. Holding this benefit constant

isolates the effect of architecture-specific investment and operating costs. The avoided-cost benefit was assumed to grow 10% annually, while architecture-specific OpEx grew 5% annually. A 10% residual-value ratio was used only to define straight-line depreciation over five years, a 10% effective tax factor was applied to EBIT, and the discount rate was 10%. For each architecture, annual depreciation was $D = (\text{CapEx} - 0.10 \times \text{CapEx})/5$ and annual net cash flow was $\text{NCF}_t = (\text{B}_t - \text{O}_t - D)(1 - 0.10) + D$. No terminal residual cash inflow was included, preserving the original five-year screening convention. NPV was the discounted sum of annual net cash flows less initial investment; IRR was the discount rate at which NPV equaled zero; and PI divided the present value of operating inflows by initial investment. Payback was recalculated from cumulative undiscounted cash flows rather than from initial investment divided by average annual cash flow. If recovery occurred during year y , $\text{payback} = (y-1) + \text{unrecovered investment at the end of year } y-1 \text{ divided by } \text{NCF}_y$. This correction is material when annual cash flows vary. The model remains a feasibility scenario rather than a market-price or realized-savings guarantee (Abdulle et al., 2022).

Regulatory readiness was assessed by mapping each architecture's data flow to Indonesia's Law No. 27 of 2022 on Personal Data Protection. The mapping covered biometric-data classification, lawful processing, transparency, high-risk impact assessment, security, retention and destruction, data-protection functions, processor accountability, and cross-border transfers. This stage did not audit an operating organization's notices, legal basis, data-protection impact assessment, processor contracts, technical logs, retention evidence, or cross-border transfer documentation. Consequently, the analysis evaluates whether an architecture can support implementation of identifiable controls; it does not demonstrate legal compliance. The final decision rule was therefore integrative and conditional: an architecture could be recommended as a deployment baseline only when its technical trade-off was supportable, its comparative financial scenario was viable, and its data flow was compatible with implementing the required governance controls.

3. RESULT AND DISCUSSION

3.1 Technical Feasibility

The simulation showed that no architecture dominated all technical metrics. Figure 1 presents the complete trajectory for processing time, failed-task rate, and total latency. Edge preserved the lowest network latency, increasing only from 0.0191 s at 100 devices to 0.0204 s at 500 devices. Its processing time, however, rose from 1.9140 to 2.8655 s, while failed tasks increased from 1.259% to 27.797%. The sharp rise from 3.314% at 200 devices to 12.021% at 300 devices, followed by a processing-time plateau near 2.86 s, indicated local compute saturation rather than a network bottleneck.

Cloud exhibited the opposite profile. Processing time remained approximately 0.175–0.181 s across all workloads, confirming ample centralized compute capacity. Total latency rose from 0.4979 to 0.9396 s, and the failed-task rate jumped from 0.382% at 300 devices to 2.839% at 400 devices and 22.137% at 500 devices. Because compute time stayed stable while both latency and task loss deteriorated, the limiting factor was consistent with WAN congestion or queue pressure. This pattern agrees with research describing bandwidth and transmission as central constraints in cloud video analytics (Badidi et al., 2023; Nguyen & Choi, 2020).

Hybrid occupied the middle ground. Processing time increased from 1.4422 to 2.2066 s, total latency remained between 0.1141 and 0.1355 s, and failed tasks grew from 0.919% to 16.736%. At 500 devices, its failure rate was 24.4% lower than Cloud and 39.8% lower than Edge, while its latency was 85.6% lower than Cloud. Hybrid nevertheless exceeded the 2 s processing-time target at 400 and 500 devices. The result supports collaborative inference findings: dividing the pipeline can reduce the weaknesses of isolated edge or cloud execution, but orchestration does not remove resource limits under sustained load (Koubaa et al., 2022; Oroceo et al., 2022).

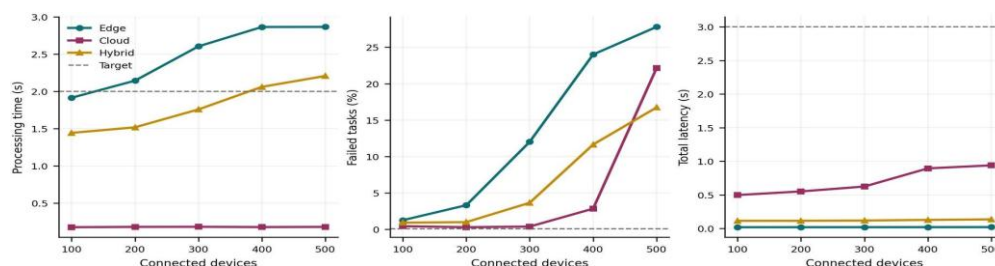


Figure 1. Technical Performance Across Connected-Device Loads

Table 3. Technical Results at 500 Connected Devices

Architecture	Processing time (s)	Failed tasks (%)	Total latency (s)	Primary bottleneck
Edge	2.8655	27.797	0.0204	Local compute saturation
Cloud	0.1792	22.137	0.9396	WAN/queue pressure
Hybrid	2.2066	16.736	0.1355	Distributed compute at high load

Table 3 compares the architectures at the maximum tested load. Under the original screening gates, all three satisfied the 3 s total-latency limit. Cloud alone satisfied the 2 s processing-time limit across all loads; Edge satisfied it only at 100 devices, and Hybrid through 300 devices. More importantly, none met the 0.1% failed-task gate even at the lowest load. This does not establish that the architectures violate a universal face-recognition SLA; rather, it shows that the chosen reliability gate is more demanding than the simulated capacity. The appropriate technical conclusion is therefore comparative rather than compliant/non-compliant: Hybrid produced the strongest relative balance at high load, but production deployment requires capacity adjustment and empirical calibration to the selected detector, embedding model, camera resolution, frame rate, and network environment.

Table 4. Threshold Sensitivity at 500 Connected Devices

Dimension	Thresholds tested	Architectures passing at 500 devices	Interpretation
Processing time	≤ 1.5 ; ≤ 2.0 ; ≤ 2.5 ; ≤ 3.0 s	1.5/2.0: Cloud; 2.5: Cloud + Hybrid; 3.0: all	Hybrid remains the middle option; the ranking is not tied only to the 2 s gate
Total latency	≤ 0.2 ; ≤ 0.5 ; ≤ 1.0 ; ≤ 3.0 s	0.2/0.5: Edge + Hybrid; 1.0/3.0: all	Cloud is most sensitive to a sub-second network bound at maximum load
Failed tasks	≤ 0.1 ; ≤ 1 ; ≤ 5 ; ≤ 10 ; ≤ 20 ; ≤ 25 ; $\leq 30\%$	≤ 10 : none; 20: Hybrid; 25: Cloud + Hybrid; 30: all	Task reliability is the dominant scaling constraint in the simulated workload

The sensitivity analysis confirms that the initial thresholds should be interpreted as screening gates rather than universal acceptance criteria. Relaxing the processing threshold from 2.0 to 2.5 s allows Hybrid to pass at 500 devices, while a 1.0 s latency bound allows all three architectures to pass. In contrast, no architecture satisfies a failed-task threshold of 10% or lower at 500 devices. Hybrid becomes the first architecture to pass only when the failure tolerance reaches 20%. Thus, the main high-load constraint is simulated task completion rather than network latency, and the recommendation for Hybrid rests on relative balance rather than on a claim of universal technical compliance.

3.2 Techno-Economic Feasibility

The economic comparison was recalculated on an architecture-consistent basis. Edge required IDR 1.98 billion CapEx and IDR 700 million first-year OpEx, reflecting localized GPU servers, power, maintenance, and personnel. Cloud reduced CapEx to IDR 790 million but raised first-year OpEx to IDR 800 million because centralized inference depended on cloud GPU usage and dedicated connectivity. Hybrid required IDR 1.50 billion CapEx and IDR 300 million first-year OpEx. Unlike the earlier one-architecture investment test, the revised analysis applies the same Year-1 avoided operating-cost benefit of IDR 850 million, the same 10% annual benefit growth, and the same financial conventions to all three architectures. This design allows economic preference to be attributed to differences in deployment cost rather than to unequal benefit assumptions. Table 5 reports the comparable five-year investment indicators. Because the modeled benefit is held constant across architectures, these values answer a narrower and more defensible question: which architecture creates the strongest financial case under the same assumed service benefit? They remain scenario estimates and should not be interpreted as realized enterprise savings or market-price guarantees.

Table 5. Comparative Economic Results under a Common Avoided-Cost Scenario

Indicator	Edge	Cloud	Hybrid
Initial CapEx	IDR 1.98 bn	IDR 0.79 bn	IDR 1.50 bn
First-year OpEx	IDR 0.70 bn	IDR 0.80 bn	IDR 0.30 bn
Year-1 avoided-cost benefit (common)	IDR 0.85 bn	IDR 0.85 bn	IDR 0.85 bn
NPV at 10%	−IDR 982.50 mn	−IDR 247.25 mn	IDR 958.96 mn
IRR	−10.04%	−0.94%	30.73%
Profitability index (PI)	0.50	0.69	1.64
Actual cumulative payback	> 5 years	> 5 years	2.60 years

Note. Common scenario assumptions: Year-1 avoided operating-cost benefit = IDR 850 million for all architectures; benefit growth = 10%/year; architecture-specific OpEx growth = 5%/year; discount rate = 10%; effective tax factor = 10%; straight-line depreciation uses a 10% residual-value ratio; no terminal residual cash inflow is included. Values are scenario outputs, not realized savings.

Under the common-benefit model, Edge generated an NPV of –IDR 982.50 million, IRR of –10.04%, and PI of 0.50; Cloud generated an NPV of –IDR 247.25 million, IRR of –0.94%, and PI of 0.69. Neither recovered its initial investment within the five-year horizon. Hybrid generated an NPV of IDR 958.96 million, IRR of 30.73%, and PI of 1.64. The payback period was recalculated from cumulative undiscounted cash flow rather than from average annual cash flow. Hybrid remained unrecovered at the end of Year 2 (–IDR 393.00 million) and generated IDR 654.975 million net cash flow in Year 3; interpolation therefore gives $2 + 393/654.975 = 2.60$ years. This replaces the earlier 2.26-year simple-payback estimate. A one-way sensitivity check also showed that Hybrid remained positive when the common Year-1 avoided-cost benefit was reduced by 20% to IDR 680 million (NPV = IDR 263.51 million; payback = 3.46 years) and when the discount rate increased to 15% (NPV = IDR 652.90 million). Cloud became NPV-positive only when the common benefit increased by approximately 10% in the tested scenario, while Edge remained NPV-negative even at a 20% benefit increase (NPV = –IDR 287.04 million). These results support Hybrid as economically preferable within the stated scenario, not as a universal investment conclusion.

3.3 Regulatory Readiness Assessment

Regulatory mapping showed why technical data flow matters, but it was treated as a readiness assessment rather than evidence of compliance. Indonesia's Personal Data Protection Law classifies biometric data as specific personal data and requires additional safeguards for high-risk processing, including large-scale or systematic monitoring and the use of new technology (Indonesia, 2022). Hybrid reduced the need to transmit continuous raw video across public networks because frame selection and face extraction occurred locally. This design can support data minimization and reduce the exposed payload, but transmitted facial crops or embeddings may still identify individuals and therefore remain subject to personal-data obligations. Architecture can reduce exposure; it cannot establish a lawful basis, transparency, proportionality, security effectiveness, retention compliance, or accountability on its own.

Table 6 therefore translates the law into controls that would have to be implemented and evidenced before production use. The organization would need to document purpose and lawful basis, provide clear notices, conduct and approve a data-protection impact assessment where required, assign responsible governance functions, enforce encryption and access control, maintain incident and audit records, define retention and secure deletion, govern processors, and assess cross-border transfers. The table should be read as a compliance-supporting readiness checklist. Demonstrating that an architecture can accommodate these controls is not equivalent to demonstrating that the deploying organization has actually complied with the law.

Table 6. Regulatory Readiness Controls for Hybrid Biometric Deployment

Control area	Required implementation action	Architecture-supporting implication
Lawful basis and transparency	Document purpose and lawful basis; issue clear notice; manage withdrawal when consent is used	Local preprocessing does not remove transparency or rights obligations
High-risk impact assessment	Complete and approve a data-protection impact assessment before go-live	Map raw images, crops, embeddings, logs, storage, and recipients
Security and retention	Encrypt data in transit/at rest; restrict access; log use; define retention and secure deletion	Protect both local gateways and cloud services
Governance and processors	Assign responsible function; execute processor agreements; audit vendors and incidents	Allocate controller–processor duties across the service chain
Cross-border transfer	Verify statutory transfer basis and contractual safeguards before offshore hosting	Prefer controlled data locations where safeguards are uncertain

3.4 Integrated Decision and Limitations

The integrated assessment supported the hypothesis conditionally. Technically, Hybrid offered the most balanced high-load trade-off, but the threshold sensitivity analysis showed that this conclusion is relative and that failed-task performance remains the dominant scaling limitation. Economically, Hybrid was the only architecture with positive NPV and PI above 1 under the same modeled avoided-cost benefit, whereas Edge and Cloud remained negative in the baseline five-year

scenario. This comparative result is stronger than evaluating Hybrid alone, but it remains conditional on the avoided-cost assumption and procurement inputs. From a regulatory perspective, Hybrid's local preprocessing can support data minimization, yet the architecture is only control-enabling and does not constitute proof of PDP-law compliance. The resulting recommendation is therefore operationally specific: use Hybrid as the pilot baseline, calibrate the simulation against the selected production face-detection and recognition pipeline, derive final thresholds from the application owner's SLA, monitor processing time and task loss, add dynamic orchestration or capacity before expansion, and retain a manual fallback for consequential access decisions.

Several limitations constrain generalization. First, EdgeCloudSim approximated infrastructure and network behavior and was structurally mapped to published face-recognition pipelines, but it was not numerically calibrated against measured per-frame execution times from the exact detector, embedding model, camera resolution, frame rate, and hardware intended for production. The technical results should therefore be interpreted as architecture-level stress behavior, not direct recognition-inference benchmarks. Second, the 2 s processing, 3 s latency, and 0.1% failed-task limits are scenario-specific screening gates rather than an organization-approved SLA; the added sensitivity analysis reduces, but does not eliminate, this assumption. Third, the study did not remeasure recognition accuracy, presentation-attack robustness, or demographic parity, even though those factors remain essential for responsible face-recognition operation (Cavazos et al., 2021; Khalil et al., 2020). Fourth, the economic benefit is a modeled avoided-cost scenario rather than an observed saving; hardware prices, cloud tariffs, tax treatment, utilization, and realized labor or administrative savings will vary by organization. Finally, the regulatory analysis identifies control requirements and architecture implications but does not audit organizational compliance and is not legal advice. A production decision should therefore combine field benchmarking, organization-specific financial data, a formal SLA, and documented legal/governance review.

4. CONCLUSION

An integrated assessment of technical, economic, and regulatory-readiness evidence identifies Hybrid as the most balanced conditional baseline for enterprise face-recognition-as-a-service under the stated assumptions. At 500 devices, Hybrid reduced peak failed tasks relative to both Edge and Cloud and avoided Cloud's higher WAN latency, although it exceeded the 2 s processing screen above 300 devices and no architecture met the stringent 0.1% failed-task gate. Threshold sensitivity confirms that task completion, rather than network latency alone, is the principal scaling constraint. The revised comparative financial model applies the same avoided operating-cost benefit to all architectures: Edge and Cloud produced negative NPVs of IDR 982.50 million and IDR 247.25 million, respectively, while Hybrid produced an NPV of IDR 958.96 million, IRR of 30.73%, PI of 1.64, and an actual cumulative payback period of 2.60 years. Hybrid's local preprocessing can also support data minimization, but this is a regulatory-readiness advantage rather than evidence of legal compliance. Resubmission conclusions are therefore deliberately conditional: a production implementation should proceed through a controlled pilot, benchmark the selected face detector and recognition/embedding pipeline on actual hardware and camera workloads, derive final service thresholds from the enterprise SLA, validate recognition accuracy and fairness, document PDP-law controls, and refresh the economic model with organization-specific avoided-cost and procurement data. Future research should perform direct pipeline calibration and multi-run simulations under fluctuating bandwidth, adaptive routing, varied camera resolutions, and measured production workloads.

ACKNOWLEDGEMENTS

The author acknowledges the School of Electrical Engineering, Telkom University, for the academic environment in which the original master's thesis was developed.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY

The simulation configuration, output tables, and calculation data supporting this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Pradipa Catya Wisesa: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Simulation; Visualization; Writing—Original Draft; Writing—Review & Editing.

REFERENCES

- Abdulle, A. A., Ali, A. F., & Abdullah, R. H. (2022). Cost-benefit analysis of public cloud versus in-house computing. *International Journal of Engineering Trends and Technology*, 70(6), 300–307. <https://doi.org/10.14445/22315381/IJETT-V70I6P231>
- Badidi, E., Moumane, K., & El Ghazi, F. (2023). Opportunities, applications, and challenges of edge-AI enabled video analytics in smart cities: A systematic review. *IEEE Access*, 11, 80543–80572. <https://doi.org/10.1109/ACCESS.2023.3300658>
- Bhatia, R., & Mathur, M. (2026). Facial recognition technology acceptance: A systematic review, cognitive implications, and automation-bias framework. *Discover Artificial Intelligence*, 6, 499. <https://doi.org/10.1007/s44163-026-01030-8>
- Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2021). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1), 101–111. <https://doi.org/10.1109/TBIOM.2020.3027269>
- Díaz Martínez, M. A., Román Salinas, R. V., Ruíz Hernández, S., Ruíz Domínguez, H. S., Cervantes Zubirías, G., & Morales Rodríguez, M. A. (2024). Artificial intelligence an essential factor for the benefit of companies: Systematic review of the literature. *Cogent Engineering*, 11(1), 2380344. <https://doi.org/10.1080/23311916.2024.2380344>
- EL Fadel, N. (2025). Facial recognition algorithms: A systematic literature review. *Journal of Imaging*, 11(2), 58.
- Hidayat, F., Elviani, U., Situmorang, G. B. G., Ramadhan, M. Z., Alunjati, F. A., & Sucipto, R. F. (2024). Face recognition for automatic border control: A systematic literature review. *IEEE Access*, 12, 37288–37309. <https://doi.org/10.1109/ACCESS.2024.3373264>
- Indonesia, R. of. (2022). *Law No. 27 of 2022 on Personal Data Protection*. https://jdih.komdigi.go.id/produk_hukum/view/id/832
- Jacob, C., Brasier, N., Laurenzi, E., Heuss, S., Mougiakakou, S.-G., Cöltekin, A., & Peter, M. K. (2025). AI for IMPACTS framework for evaluating the long-term real-world impacts of AI-powered clinician tools: Systematic review and narrative synthesis. *Journal of Medical Internet Research*, 27, 67485. <https://doi.org/10.2196/67485>
- Khalil, A., Ahmed, S. G., Khattak, A. M., & Al-Qirim, N. (2020). Investigating bias in facial analysis systems: A systematic review. *IEEE Access*, 8, 130751–130761. <https://doi.org/https://doi.org/10.1109/ACCESS.2020.3006051>
- Koubaa, A., Ammar, A., Kanhouc, A., & AlHabashi, Y. (2022). Cloud versus edge deployment strategies of real-time face recognition inference. *IEEE Transactions on Network Science and Engineering*, 9(1), 143–160. <https://doi.org/https://doi.org/10.1109/TNSE.2021.3055835>
- Merenda, M., Porcaro, C., & Iero, D. (2020). *Edge machine learning for AI-enabled IoT devices: A review*. *Sensors*, 20(9), 2533. <https://doi.org/https://doi.org/10.3390/s20092533>
- Nguyen, D. V., & Choi, J. (2020). Toward scalable video analytics using compressed-domain features at the edge. *Applied Sciences*, 10(18), 6391. <https://doi.org/https://doi.org/10.3390/app10186391>
- Oroceo, P. P., Kim, J.-I., Caliwag, E. M. F., Kim, S.-H., & Lim, W. (2022). Optimizing face recognition inference with a collaborative edge–cloud network. *Sensors*, 22(21), 8371. <https://doi.org/https://doi.org/10.3390/s22218371>
- Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., & Dai, H. (2018). A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Communications Surveys & Tutorials*, 20(4), 3098–3130. <https://doi.org/https://doi.org/10.1109/COMST.2018.2841349>
- Phu, G. N., Thi, T. H., & Bich, H. T. N. (2025). The impact of cloud computing technology on cloud accounting adoption and financial management of businesses. *Humanities and Social Sciences Communications*, 12, 851. <https://doi.org/https://doi.org/10.1057/s41599-025-05190-3>
- Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30–39. <https://doi.org/https://doi.org/10.1109/MC.2017.9>
- Sonmez, C., Ozgovde, A., & Ersoy, C. (2018). EdgeCloudSim: An environment for performance evaluation of edge computing systems. *Transactions on Emerging Telecommunications Technologies*, 29(11), 3493. <https://doi.org/https://doi.org/10.1002/ett.3493>

- Ssegawa, J. K., & Muzinda, M. (2021). Feasibility assessment framework (FAF): A systematic and objective approach for assessing the viability of a project. *Procedia Computer Science*, 181, 377–385. <https://doi.org/https://doi.org/10.1016/j.procs.2021.01.180>
- Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D. (2025). Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics*, 13(11). <https://doi.org/https://doi.org/10.3390/math13111878>
- Wang, Y., Wang, W., Liu, D., Jin, X., Jiang, J., & Chen, K. (2023). Enabling edge-cloud video analytics for robotics applications. *IEEE Transactions on Cloud Computing*, 11(2), 1500–1513. <https://doi.org/https://doi.org/10.1109/TCC.2022.3142066>
- Xu, R., Razavi, S., & Zheng, R. (2023). Edge video analytics: A survey on applications, systems and enabling techniques. *IEEE Communications Surveys & Tutorials*, 25(4), 2951–2982. <https://doi.org/https://doi.org/10.1109/COMST.2023.33230>