Journal of Integrated Science, Technology and Management, 2025; 1(2):22-29

Journal of Integrated Science, Technology and Management

Research Article

A Hybrid Speech–Text Foundation Model for Noise-Robust Multilingual Automatic Speech Recognition

Yusupov Shermatilla[1*], Kratika Jain[2]

[1]*Department of Philosophy, Tashkent State Transport University, Tashkent, Uzbekistan*
[2]*Researcher*

ARTICLE INFO

Article history:
Received: 06-02-2025
Revised: 14-06-2025
Accepted: 23-08-2025
Published: 18-11-2025

Keywords:
Automatic speech recognition, multilingual speech, foundation models, noise robustness, cross-modal attention, speech–text fusion, code-switching, low-resource languages.

DOI: 10.66592/xjc846te

ABSTRACT

Automatic speech recognition (ASR) in multilingual settings remains vulnerable to nonstationary noise, code-switching, accented pronunciation, limited labeled data, and ambiguity among acoustically similar words. This paper proposes a hybrid speech–text foundation model integrating a pretrained multilingual acoustic encoder, a noise-aware feature adaptation module, a multilingual text encoder, and confidence-gated cross-modal decoding. The system combines acoustic likelihood with contextual linguistic evidence while controlling semantic overcorrection through uncertainty-aware fusion. A reproducible evaluation protocol is specified for clean and noise-corrupted multilingual corpora, including low-resource languages and mixed-language utterances. Word error rate (WER), character error rate (CER), language identification accuracy, real-time factor, and calibration are proposed as evaluation measures. Illustrative numerical scenarios are provided solely to demonstrate the planned comparative and ablation analyses; they are not measured experimental outcomes. The proposed architecture offers a testable research direction for resilient multilingual transcription under practical deployment constraints.

INTRODUCTION

Automatic speech recognition converts spoken utterances into textual sequences and is central to accessibility tools, conversational agents, meeting transcription, public-service interfaces, and voice-operated industrial systems. Although recognition quality has improved substantially, accuracy is not uniform across languages, acoustic environments, or speaker populations. Many deployed systems perform well for standardized speech recorded with high-quality microphones but degrade when faced with reverberation, traffic noise, overlapping voices, regional accents, and spontaneous code-switching. These conditions frequently occur together, making multilingual recognition a compound rather than an isolated modeling problem.

A multilingual speech signal carries information at several levels. The waveform encodes phonetic events, speaking rate, intonation, microphone response, and environmental interference. The transcript reflects a language-dependent writing system, lexical choices, grammatical conventions, and domain-specific meaning. Recognition models that focus exclusively on the acoustic pathway may struggle to resolve competing hypotheses when noise masks short consonants or weak syllables. Conversely, language-heavy decoders may produce fluent but unsupported words when the speech signal is uncertain. Reliable transcription therefore requires an explicit balance between acoustic evidence and textual plausibility.

Traditional ASR pipelines separated feature extraction, acoustic modeling, pronunciation lexicons, and language modeling. Such systems offered modularity but required considerable linguistic engineering and often failed to generalize to new languages. End-to-end neural approaches reduced these dependencies by learning direct mappings from audio to token sequences. Connectionist temporal classification, attention-based encoder–decoder systems, and transducer models each provide mechanisms for aligning variable-length speech with textual outputs. Their effectiveness nevertheless depends on suitable training

***Corresponding Author:** Yusupov Shermatilla
Address: Department of Philosophy, Tashkent State Transport University, Tashkent, Uzbekistan
Email ✉ yusupov_she@tstu.uz **ORCID:** 0009-0001-3684-0066

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, Yusupov Shermatilla and Kratika Jain, 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.

distributions and sufficient coverage of pronunciation and environmental variability.

Self-supervised pretraining transformed this landscape by allowing speech encoders to learn useful representations from large collections of unlabeled recordings. Multilingual pretraining further encourages shared acoustic structure across languages, permitting adaptation when labeled examples are scarce. However, language-independent features alone cannot fully resolve orthographic ambiguity, rare named entities, or rapidly changing terminology. Speech–text integration is attractive because written material is often easier to obtain than accurately aligned multilingual speech recordings, although text corpora also introduce their own biases and quality concerns.

Noise robustness is especially challenging because real acoustic interference differs from idealized additive white noise. A field recording may contain transient horns, machinery, music, reverberant reflections, competing speech, packet loss, or clipping. Noise may alter both spectral characteristics and temporal boundaries, disrupting attention alignment and token prediction. Models trained only on clean speech can learn fragile cues that disappear at low signal-to-noise ratios. Robust evaluation must therefore include several noise classes and severity levels, along with unseen acoustic conditions rather than only familiar augmentation patterns.

Language diversity creates another difficulty. High-resource languages benefit from large transcribed corpora, standardized text normalization, and mature evaluation conventions. Low-resource languages may have multiple scripts, inconsistent spelling, dialectal variation, or limited digital text. Transcription can also involve mixed-script utterances and borrowed terms. In code-switched speech, a single sentence may contain lexical items from two or more languages, and an utterance-level language identifier can be insufficient. The recognizer must preserve local language transitions without forcing all words into a single vocabulary.

Foundation models offer a useful starting point because pretraining across tasks and languages produces reusable representations. Yet a foundation model is not automatically robust, fair, or computationally efficient. Model scale can increase inference cost and memory requirements, which matters for mobile devices, rural connectivity, and latency-sensitive applications. A practical design should permit parameter-efficient adaptation, selective use of expensive text reasoning, and predictable decoding time. It should also distinguish genuine improvements in acoustic recognition from gains caused by aggressive linguistic rewriting.

The central research hypothesis in this study is that uncertainty-aware fusion of multilingual acoustic representations and textual context can improve recognition under noise while avoiding unsupported corrections. The proposed hybrid model first estimates acoustic tokens and language cues, then retrieves contextual information from a multilingual text encoder. A learned gate weighs the reliability of each stream using acoustic confidence, noise severity, and language uncertainty. When the audio evidence is strong, the gate prioritizes faithful transcription; when it is weak, textual evidence can assist in ranking plausible alternatives without freely generating unrelated content.

This design differs from simple post-processing, in which a language model rewrites an entire ASR transcript after decoding. Instead, cross-modal information is introduced within hypothesis scoring and constrained by the acoustic lattice or candidate set. Such constraints are important for names, numerical expressions, and specialized vocabulary, where a fluent correction may be factually incorrect. The system also employs augmentation and consistency regularization to reduce sensitivity to perturbations, and language-balanced sampling to avoid allowing dominant languages to overwhelm smaller ones.

The contributions of this paper are a modular hybrid speech–text architecture, a confidence-gated fusion mechanism, a combined training objective, and an evaluation framework spanning multilingual accuracy, noise robustness, code-switching, efficiency, and calibration. The study additionally provides an ablation plan that can isolate the effect of acoustic pretraining, denoising, text context, and gating. Because actual model training has not been conducted for this manuscript, all proposed performance figures are illustrative and are not represented as empirical discoveries. The resulting document is a research design and publication-ready methodological draft pending real experimentation.

Related Work

Large-vocabulary speech recognition initially relied on probabilistic acoustic models, pronunciation dictionaries, and statistical language models. Hidden Markov model systems captured temporal variation through state transitions, while Gaussian mixtures and later deep neural networks estimated acoustic likelihoods. Their modular construction facilitated targeted optimization, but it also required language-specific phonetic resources and complicated engineering. End-to-end systems subsequently simplified the pipeline by optimizing transcript likelihood directly, although sequence alignment and decoding remained major challenges.

Connectionist temporal classification (CTC), introduced by Graves and colleagues, addressed unsegmented sequence labeling through a blank symbol and a sum over valid alignments. Its conditional independence assumptions support efficient inference but can limit contextual modeling. Attention-based sequence-to-sequence systems model dependencies among output tokens more explicitly, at the cost of potentially unstable alignments for long recordings. Recurrent neural network transducers combine streaming alignment and autoregressive prediction and have become important for low-latency recognition. These approaches provide complementary design principles for the proposed hybrid decoder.

Baevski et al. demonstrated that wav2vec 2.0 can learn strong acoustic representations through contrastive self-supervision on raw audio. Its masked latent modeling reduces dependence on expensive manual transcription and supports fine-tuning for speech recognition. Cross-lingual extensions explored the possibility that phonetic representations transfer across languages. Conneau et al. reported multilingual self-supervised speech representations that improved low-resource adaptation, supporting the argument that acoustic knowledge need not be learned independently for every language.

XLS-R expanded multilingual representation learning through large-scale pretraining across 128 languages. Babu et al. showed that increasing the linguistic breadth and scale of unlabeled audio can improve speech recognition and related speech tasks. The approach is particularly relevant when labeled examples are limited but raw speech can be collected. However, broad pretraining does not guarantee consistent performance for every dialect or noise type, and a model trained on uneven language distributions may still show systematic gaps across communities.

Radford et al. introduced Whisper as a large-scale weakly supervised approach trained on hundreds of thousands of hours of multilingual and multitask audio. The work demonstrated strong transfer to diverse recognition conditions and helped establish large pretrained encoder–decoder models as general-purpose speech systems. Its success highlights the value of heterogeneous training data, but weak supervision can also contain transcription noise, mismatched segments, and language imbalances. Consequently, robust deployment still requires domain-specific validation rather than assuming universal performance.

Multilingual benchmarks have become crucial for measuring generalization beyond a few well-resourced languages. FLEURS, proposed by Conneau et al., provides parallel speech material across 102 languages and supports recognition and language identification research. Such benchmarks make it possible to compare performance across language families and writing systems. Nevertheless, benchmark speech is not a complete substitute for uncontrolled recordings, spontaneous conversations, code-switching, or noisy public environments. Evaluations should combine established corpora with carefully documented perturbations and, where permitted, natural field audio.

The role of textual language modeling in ASR predates neural foundation models. N-gram language models, recurrent language models, and Transformer language models have been used to rescore acoustic hypotheses. Shallow fusion interpolates language-model scores during decoding, while rescoring revisits an N-best list after an initial pass. These methods can reduce lexical errors when contextual information is accurate, but overly strong text priors may override genuine acoustic evidence. The resulting hallucination risk is particularly serious when the input is silent, heavily corrupted, or contains unfamiliar proper nouns.

Cross-modal learning attempts to connect speech representations with text semantics more directly. Shared embeddings, contrastive alignment, speech-to-text pretraining, and multimodal sequence modeling can help establish correspondences between acoustically different expressions and related linguistic meanings. This connection is promising for multilingual recognition because semantically similar phrases may appear across different languages. However, semantic similarity is not equivalent to transcript correctness: the goal of ASR is to reproduce what was spoken, not to paraphrase its intended meaning. Therefore, cross-modal objectives require transcript-preserving constraints.

Noise-robust ASR research commonly uses speed perturbation, reverberation simulation, spectral masking, additive environmental noise, and speech enhancement. SpecAugment demonstrated the utility of masking time and frequency regions during acoustic training. Enhancement front ends can improve intelligibility in some conditions, but artifacts introduced by denoisers may remove weak phonetic cues. Joint training and consistency objectives can reduce this mismatch by encouraging the recognizer to produce stable transcripts for clean and perturbed versions of the same utterance.

Code-switching and low-resource evaluation expose weaknesses in conventional language conditioning. Systems that select one language for an entire utterance may mishandle local transitions, while large shared tokenizers can underrepresent less common scripts. Balanced sampling, language-aware adapters, character-level error reporting, and segment-level language identification offer possible remedies. Recent multilingual spoken-language-understanding work also emphasizes that accurate recognition and semantic interpretation are related but distinct capabilities. This paper builds on these findings by treating contextual text as a calibrated auxiliary signal, not as a replacement for acoustic grounding.

PROPOSED METHODOLOGY

Problem formulation and architecture

Let $x = (x_1,\ldots,xT)$ be a noisy waveform, l a language or language-mixture label, c optional domain context, and $y = (y_1,\ldots,yU)$ the target transcript. The goal is to estimate $y^* = \text{argmax_y } P(y \mid x,c)$ under constraints on error rate, latency, and unsupported text generation. The architecture contains (i) waveform normalization and augmentation, (ii) a multilingual speech foundation encoder, (iii) noise-aware adapters, (iv) a multilingual text-context encoder, (v) confidence-gated cross-modal fusion, and (vi) constrained beam decoding.

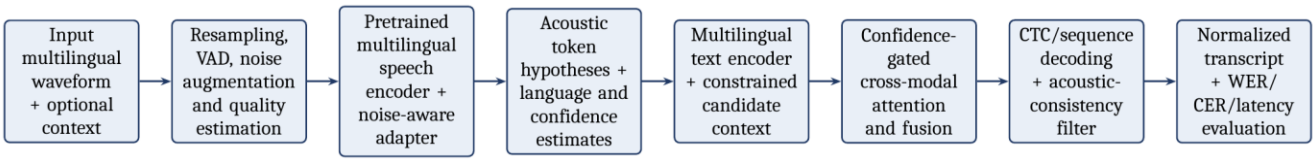


Figure 1: Proposed end-to-end workflow of the hybrid speech–text recognition system.

Acoustic representation and noise adaptation

A multilingual encoder $f\theta$ converts the waveform into latent acoustic features $Hs = f\theta(x)$. A lightweight noise adapter $g\varphi$ transforms these features using estimated signal quality q: $Hn = g\varphi(Hs,q)$. Noise augmentation samples a corruption operator A from a predefined family, producing $\tilde{x} = A(x;SNR,r)$, where SNR is signal-to-noise ratio and r denotes reverberation conditions. Augmentation must be applied only to training data, and evaluation corruption seeds must be held fixed for reproducibility.

$$Hs = f\theta(x),\ Hn = g\varphi(Hs, q),\ \tilde{x} = A(x; SNR, r) \quad (1)$$

Speech–text fusion

The text encoder produces $Ht = e\psi(c,h)$, where h denotes N-best acoustic hypotheses or retrieved domain vocabulary. Cross-attention uses acoustic states as queries and text states as keys and values. A confidence gate $\alpha t \in [0,1]$ depends on acoustic entropy, estimated noise severity, and language uncertainty. This gate prevents context from dominating when acoustic evidence is reliable.

$$Ct = \text{softmax}[(HnWQ)(HtWK)^T / \sqrt{d}]\ (HtWV) \quad (2)$$

$$\alpha t = \sigma(w^T[Hn,t\ ;\ Ct\ ;\ qt\ ;\ ut] + b),\ Zt = (1-\alpha t)Hn,t + \alpha tCt \quad (3)$$

Here WQ, WK, and WV are trainable projections, d is attention dimensionality, σ is the logistic sigmoid, qt is the noise-quality descriptor, and ut measures predictive uncertainty. Fusion states Z are passed to a transcript decoder, with an acoustic-consistency penalty on candidate sequences that lack sufficient support in the speech model.

Training objectives

$$L = \lambda 1\ LCTC + \lambda 2\ Lseq + \lambda 3\ Lcontrast + \lambda 4\ Lcons + \lambda 5\ Llang + \lambda 6\ Lcal \quad (4)$$

LCTC is the alignment-based transcription loss; Lseq is teacher-forced token cross-entropy; Lcontrast aligns paired speech and text embeddings using positive and in-batch negative pairs; Lcons encourages stable predictions across clean and augmented recordings; Llang supervises language identification; and Lcal penalizes overconfident incorrect predictions. Coefficients λi are selected on a validation set, not on held-out test data. When paired text is unavailable, the contrastive term is disabled for that example.

$$Lcons = DKL[P\theta(\cdot|x)\ ||\ P\theta(\cdot|\tilde{x})] \quad (5)$$

Decoding, data protocol, and evaluation

During inference, a beam search scores candidate transcripts using acoustic, sequence, and text-context terms. The text weight is adaptively reduced when evidence for the candidate is inconsistent with the audio. Silence detection and minimum acoustic-support thresholds are applied to reduce fabricated words. The proposed study should use speaker-disjoint train/development/test partitions and language-balanced batches. Publicly documented sources such as FLEURS and Common Voice can support multilingual evaluation; noise may be added at 20, 10, 5, and 0 dB SNR with separate stationary and nonstationary categories. Real recordings should supplement synthetic noise when licensing and consent permit.

$$Score(y) = \log P_{acoustic}(y|x) + \beta(x) \log P_{text}(y|c) - \gamma\ Halluc(y,x) \quad (6)$$

$$WER = 100(S + D + I)/N;\ CER = 100(Sc + Dc + Ic)/Nc \quad (7)$$

S, D, and I denote word substitutions, deletions, and insertions; N is the number of reference words. The character counterparts are indexed by c. Report macro-averaged language WER in addition to pooled WER to avoid obscuring weak-language performance. Real-time factor (RTF) is inference duration divided by audio duration; smaller values indicate faster processing. Statistical uncertainty should be assessed with paired bootstrap resampling by utterance and repeated training seeds.

Results and Discussion: Illustrative Scenarios

IMPORTANT: The following values are hypothetical, internally consistent example outputs intended to illustrate the proposed reporting format. No training run, benchmark evaluation, or statistical test was performed for this manuscript. These values must not be presented as observed findings or cited as evidence of model superiority.

Illustrative model comparison

Table 1: Hypothetical model comparison. Lower WER, CER, and RTF are preferable; accuracy and computational cost must be assessed together.

System	*Clean WER (%)*	*10 dB WER (%)*	*0 dB WER (%)*	*CER (%)*	*RTF*
CTC baseline	15.8	24.7	40.5	10.2	0.19
XLS-R fine-tuned	12.6	20.4	34.1	8.4	0.31
Speech encoder + text rescoring	11.9	18.8	31.6	7.9	0.39
Proposed hybrid (illustrative)	10.7	16.3	27.9	7.0	0.43

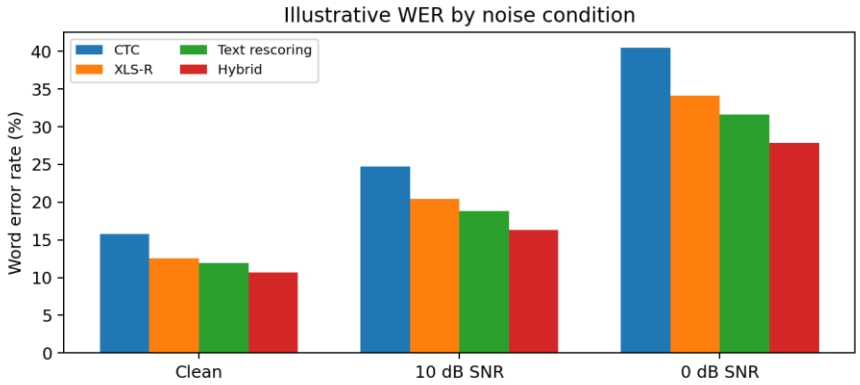


Figure 2: Hypothetical noise-condition comparison; plotted values are not measured results.

In this illustrative scenario, the hybrid system exhibits a lower error rate at every evaluated noise level, with a larger absolute advantage under severe corruption. Such a pattern would support the hypothesis that contextual fusion is most helpful when acoustic information is unreliable. However, this inference is conditional: real experiments must determine whether the improvement persists across speakers, languages, unseen noise, and multiple training seeds. The higher RTF in the example also illustrates a practical trade-off between recognition accuracy and decoding cost.

Illustrative language-level analysis

Table 2: Illustrative language-wise errors, not benchmark measurements.

Language condition	*Acoustic-only WER (%)*	*Hybrid WER (%)*	*Absolute difference (pp)*
English	12.0	10.4	1.6
Hindi	19.1	15.5	3.6
Spanish	14.2	11.9	2.3
Arabic	21.0	17.6	3.4
Code-switched Hindi–English	28.5	22.8	5.7

The illustrative code-switching gap suggests a useful target for experimental investigation: language-aware context might be especially valuable when word boundaries and vocabulary choices shift within an utterance. A real study should include segment-level language labels and evaluate transliteration and punctuation separately. Macro-averaged results, speaker breakdowns, and confidence intervals are necessary to distinguish robust cross-lingual gains from dataset composition effects.

Illustrative ablation study

Table 3: Hypothetical ablation results designed to show the required experimental analysis.

Configuration	*10 dB WER (%)*	*0 dB WER (%)*
Full hybrid model	16.3	27.9
Without noise-aware adapter	18.2	31.0
Without text-context encoder	19.0	32.4
Without confidence gate	17.8	30.6
Without consistency regularization	17.4	29.8

Ablation testing should hold training data, compute budget, decoding parameters, and random seeds as constant as feasible. If removing the text encoder increases WER but decreases insertion errors, that outcome would indicate a meaningful trade-off requiring further calibration. In addition to average error, report hallucination rate on silence and nonspeech audio, proper-noun accuracy, worst-language WER, and expected calibration error. Error inspection should categorize substitutions caused by acoustic masking separately from language-model overcorrection.

Limitations and Reproducibility

The proposed system may require substantial memory and inference compute, and its performance may depend on the availability of high-quality multilingual text. Noise augmentation cannot represent all real recording conditions, while text retrieval may introduce bias or privacy concerns. The use of an external language model creates a risk of fluent but inaccurate output. Future implementation must publish model checkpoint identifiers, dataset versions and licenses, audio preprocessing, text normalization rules, hyperparameters, hardware, training seeds, inference settings, and full per-language results. Any comparisons must use matching evaluation splits and normalization conventions. The illustrative figures in this manuscript cannot support a statistical significance claim.

Conclusion

This paper presents a methodological framework for multilingual ASR that combines a pretrained speech encoder, noise-aware acoustic adaptation, multilingual textual context, and uncertainty-controlled cross-modal fusion. Its design aims to address noisy recordings, low-resource languages, code-switching, and hallucination-prone decoding while preserving acoustic fidelity. The proposed training objective integrates transcription, cross-modal alignment, perturbation consistency, language supervision, and calibration. A detailed experimental protocol specifies relevant benchmarks, noise conditions, error metrics, and ablation comparisons. Illustrative tables and a graph demonstrate how empirical results should be reported, but do not establish performance gains. The principal next step is to implement the model and evaluate it against reproducible baselines with statistical uncertainty, fairness analysis, and realistic deployment measurements.

References

1. Talasila, D. (2023). Implementation of cloud-based disaster recovery models for minimizing business downtime and ensuring operational continuity through geographically distributed backup systems. International Journal on Recent and Innovation Trends in Computing and Communication, 11(8), 867–874. https://www.ijritcc.org/index.php/ijritcc/article/view/11934
2. Anchala, S. (2024). Deployment of AI-enabled surveillance systems in urban areas for strengthening public safety and crime prevention through real-time facial recognition and anomaly detection. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(10), 16116–16125. https://doi.org/10.15680/IJMRSET.2024.0710079
3. Gupta, A. (2024). Containerization, virtualization, and cloud security: Analyzing vulnerabilities in Kubernetes, Docker, and hypervisor-based architectures with mitigation strategies. International Journal of Advanced Research in Education and Technology, 11(3), 1322–1331. https://doi.org/10.15680/IJARETY.2024.1103067
4. Dwivedi, D. (2018). DevSecOps and API security: Embedding secure development practices, automated testing, and continuous monitoring in API-centric architectures. International Journal of Innovative Research in Computer and Communication Engineering, 6(1), 511–519. https://doi.org/10.15680/IJIRCCE.2018.0601084
5. 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).
6. Gupta, A. (2021). Deployment of advanced endpoint detection in remote work environments for preventing malware intrusions and data exfiltration through behavioral analytics and AI-driven responses. NeuroQuantology, 19(9), 1206–1224. https://doi.org/10.48047/nq.2021.19.9.NQ21199
7. Bhanushali, B. (2025). Leveraging block chain and zero-knowledge proofs for enhancing privacy in financial transactions, securing sensitive data, and mitigating cyber security risks in digital payment systems. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 14(1), 130–138. https://doi.org/10.15662/IJAREEIE.2025.1401020
8. 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).
9. Dwivedi, D. (2024). Agentic AI and global governance: Addressing the challenges of regulating autonomous artificial agents across borders in a digitally interconnected world. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 13(11), 2527–2534. https://doi.org/10.15662/IJAREEIE.2024.1311011
10. Aggarwal, A. (2021). Implementation of machine learning in cloud SIEM solutions for enhancing threat detection accuracy and reducing alert fatigue through correlation-based anomaly detection.

NeuroQuantology, 19(10). https://doi.org/10.48047/nq.2021.19.10.NQ21226

11. Devulapalli, N. (2020). Implementation of compliance-as-code in regulated DevSecOps environments for automating audit readiness and reducing human error through scripted control validation. Journal of Critical Reviews, 7(1), 2651–2657. https://doi.org/10.69980/jcr.v7:i1.13405
12. S. Goyal, K. K. Gattupalli, B. Rallabandi and M. S. Medarametla, "Intent-Driven Zero-Touch Provisioning in Private 5G Networks with AI/ML Automation," 2025 UPWIECON, Dehradun, India, 2025, pp. 839–845, doi: 10.1109/UPWIECON67212.2025.11390129.
13. 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
14. 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.
15. Khan, S. (2022). Sales force security in mobile and remote work environments: Addressing authentication challenges, secure synchronization, and device-level threats. International Journal of Innovative Research in Computer and Communication Engineering, 10(12), 8735–8743. https://doi.org/10.15680/IJIRCCE.2022.1012040
16. 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.
17. Chatrath, A. (2025). Integration of AI in air quality monitoring systems for enhancing environmental health and public awareness through predictive analytics and real-time sensing networks. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 8(1). https://doi.org/10.15680/IJMRSET.2025.0801068
18. Dwivedi, D. (2025). A comparative study of public, private, and hybrid cloud security architectures: Evaluating trade-offs in scalability, privacy, cost, and risk management. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 8(1), 485–493. https://doi.org/10.15680/IJMRSET.2025.0801069
19. 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.
20. Rangappa, A. S. (2022). The use of blockchain for auditable and tamper-proof database security logs: Enhancing transparency and trust in data access monitoring. International Journal on Recent and Innovation Trends in Computing and Communication, 10(9), 254–261. https://www.ijritcc.org/index.php/ijritcc/article/view/11925
21. Jain, Arpit, Adesh Kumar, and Sanjeev Sharma. "Comparative design and analysis of mesh, torus and ring NoC." Procedia Computer Science 48 (2015): 330–337.
22. Aggarwal, A. (2024). Integration of serverless security mechanisms in Function-as-a-Service platforms for preventing execution abuse and data leakage through contextual isolation techniques. International Journal of Innovative Research in Computer and Communication Engineering, 12(5). https://doi.org/10.15680/IJIRCCE.2024.1205374
23. Ahuja, R. (2023). Implementation of AI-driven input validation in user interfaces for preventing injection attacks and ensuring data sanitation through adaptive filtering and pattern recognition. International Journal on Recent and Innovation Trends in Computing and Communication, 11(8), 884–891. https://www.ijritcc.org/index.php/ijritcc/article/view/11936
24. Devulapalli, N. (2023). Database security in disaster recovery and business continuity planning: Ensuring data availability, backup integrity, and secure failover mechanisms. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 820–827. https://www.ijritcc.org/index.php/ijritcc/article/view/11944
25. Aggarwal, A. (2020). Deployment of multi-factor authentication in cloud-based enterprise applications for preventing credential theft and unauthorized access through adaptive verification layers. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 3(8). https://doi.org/10.15680/IJMRSET.2020.0308009
26. 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.
27. Rangappa, A. S. (2024). Generative AI in human-machine collaboration: Exploring co-creation dynamics, shared agency, and the role of human oversight in AI-augmented creative and analytical tasks. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 13(4), 985–992. https://doi.org/10.15662/IJAREEIE.2024.1304077
28. Chatrath, A. (2024). The role of encryption and key management in cloud security: Balancing performance, scalability, and confidentiality in multi-tenant and distributed data storage systems. International Journal of Innovative Research in Science, Engineering and Technology, 13(7). https://doi.org/10.15680/IJIRSET.2024.1307187
29. Gupta, A. (2022). The role of artificial intelligence in transforming healthcare delivery: A critical analysis of machine learning models, predictive diagnostics, and ethical challenges in medical decision-making. NeuroQuantology, 20(3), 1383–1399. https://doi.org/10.48047/nq.2022.20.3.NQ23005
30. K. Cherladine, B. R. Rallabandi, A. Goel, K. Kaushik, A. Dhillon and M. Soni, "Generative Adversarial Defense Models for Simulated Cyberattack Scenarios in Virtual Networks," 2025 ICDISS, Faridabad, India, 2025, pp. 1–6, doi: 10.1109/ICDISS68238.2025.11320686.
31. Ahuja, R. (2024). Deployment of content security policies in front-end frameworks for blocking cross-site scripting and clickjacking through declarative access restrictions. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(11), 17234–17242. https://doi.org/10.15680/IJMRSET.2024.0711132
32. Talasila, D. (2020). Deployment of shift-left security practices in agile development for embedding security controls early in the lifecycle and reducing technical debt through dev-centric training. Turkish Online Journal of Qualitative Inquiry, 11(1), 643–651. https://doi.org/10.69980/jzqq0t93
33. Aggarwal, A. (2022). Cloud security in artificial intelligence workloads: Protecting model training, data confidentiality, and AI-generated insights in cloud-based machine learning pipelines. NeuroQuantology, 20(3), 1367–1382. https://doi.org/10.48047/nq.2022.20.3.NQ23004
34. Ahuja, R. (2025). The risks and opportunities of generative AI in cyber security: Investigating the dual-use nature of generative models for phishing, malware development, and defence through synthetic data and automated threat detection. International Journal of Innovative Research in Computer and Communication Engineering, 13(2), 1758–1767. https://doi.org/10.15680/IJIRCCE.2025.1302059
35. Devulapalli, N. (2019). Post-quantum cryptography for database security: Preparing for the impact of quantum computing on encrypted data storage and secure queries. Journal of Critical Reviews, 6(1), 614–620. https://doi.org/10.69980/jcr.v6:i1.13396
36. Anchala, S. (2023). Integration of confidential computing in multi-cloud environments for enhancing data privacy and regulatory compliance through encrypted processing and isolated execution. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 5698–5709. https://www.ijritcc.org/index.php/ijritcc/article/view/11933
37. Chatrath, A. (2021). Bias, fairness, and ethics in artificial intelligence: A comprehensive study of algorithmic discrimination, transparency challenges, and governance frameworks. International Journal of

Multidisciplinary Research in Science, Engineering and Technology, 4(1). https://doi.org/10.15680/IJMRSET.2021.0401014

38. Mahadasu, S. S. (2019). Performance-aware VNF placement for distributed 5G core networks over heterogeneous edge and cloud infrastructure. International Journal of Engineering Science and Advanced Technology, 19(10), 5–11. https://doi.org/10.64771/ijesat.2019.v19.i10.pp3849
39. S. Goyal, K. K. Gattupalli, B. Rallabandi and M. S. Medarametla, "Integrating Terrestrial and Non-Terrestrial Networks for Reliable Rural Coverage in Agriculture and Smart Grids," 2025 UPWIECON, Dehradun, India, 2025, pp. 846–850, doi: 10.1109/UPWIECON67212.2025.11390331.
40. Dwivedi, D. (2024). Deployment of AI-based CAPTCHA systems in web applications for preventing automated bot attacks and brute force attempts through adaptive human verification. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(7), 13529–13536. https://doi.org/10.15680/IJMRSET.2024.0707292
41. Bhanushali, B. (2020). Enhancing alert investigation process with feedback loop to increase efficiently and model tuning. International Journal of Innovative Research in Computer and Communication Engineering, 8(8), 3366–3373. https://doi.org/10.15680/IJIRCCE.2020.0808100
42. N. Tripathi, K. K. Gattupalli, B. Rallabandi and M. S. Medarametla, "AI-Native Cloud-RAN Orchestration for Enterprise Private 5G Using Digital Twin Models," 2025 1st IEEE UP Section WIE Int. Conf. on Electrical Electronics and Computer Engineering (UPWIECON), Dehradun, India, 2025, pp. 851–857, doi: 10.1109/UPWIECON67212.2025.11390348.
43. 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
44. Mahadasu, S. S. (2021). Telemetry-assisted mobility management for 5G standalone networks with distributed edge user-plane functions. International Journal of Engineering Research and Science & Technology, 17(1), 141–145. https://doi.org/10.62643/ijerst.2019.v17.n1.4406
45. Rangappa, A. S. (2023). An analytical study of role-based access control (RBAC) and attribute-based access control (ABAC) in database security for multi-tenant and cloud-based architectures. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 812–819. https://www.ijritcc.org/index.php/ijritcc/article/view/11943
46. Kumar, S., Jain, A., & Moparthi, N. R. (2022). Enhanced method of object tracing using extended Kalman filter via binary search algorithm.
47. Bhanushali, B. (2024). Artificial intelligence-powered anti-money laundering framework for enhancing suspicious transaction detection, automating compliance processes, and strengthening fraud prevention in financial institutions. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(10), 16126–16134. https://doi.org/10.15680/IJMRSET.2024.0710080
48. Chatrath, A. (2021). Implementation of AI in recruitment platforms for improving candidate screening and diversity inclusion through bias-free resume parsing and skill matching. International Journal of Innovative Research in Computer and Communication Engineering, 9(12). https://doi.org/10.15680/IJIRCCE.2021.0912046
49. 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.
50. Anchala, S. (2023). Utilization of blockchain technology in cloud storage systems for enhancing data integrity and preventing unauthorized modifications through decentralized validation. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5), 629–637. https://www.ijritcc.org/index.php/ijritcc/article/view/11942
51. 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.
52. K. M. Zubair, B. R. Rallabandi, A. Yadav, K. Kaushik, S. Wankhade and M. Soni, "Explainable AI-Driven Intrusion Response Architecture for Autonomous Cyber Defense Systems," 2025 Int. Conf. on Digital Innovations for Sustainable Solutions (ICDISS), Faridabad, India, 2025, pp. 1–6, doi: 10.1109/ICDISS68238.2025.11320639.
53. Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management. International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.
54. Talasila, D. (2024). Adoption of data loss prevention strategies in SaaS platforms for safeguarding confidential information and preventing unauthorized transfers through real-time content inspection. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 7(7), 13537–13545. https://doi.org/10.15680/IJMRSET.2024.0707293
55. Mahadasu, S. S. (2022). GitOps-based continuous delivery of cloud-native 5G network functions with automated configuration validation. International Journal of Data Science and IoT Management System, 1(1), 50–56. https://doi.org/10.64751/ijdim.2022.v1.n1.1286
56. Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.
57. Bhanushali, B. (2024). Tokenization of real-world assets through block chain for improving liquidity, enabling fractional ownership, and reducing transaction costs in global financial markets. International Journal of Advanced Research in Education and Technology, 11(4), 1809–1816. https://doi.org/10.15680/IJARETY.2024.1104071
58. Talasila, D. (2021). A risk-based framework for database security: Identifying, classifying, and prioritizing threats for adaptive security policy enforcement. International Journal on Recent and Innovation Trends in Computing and Communication, 9(1), 54–61. https://www.ijritcc.org/index.php/ijritcc/article/view/11919
59. Mahadasu, S. S. (2020). Cloud-native life cycle management of virtualized RAN functions using Kubernetes operators and policy-driven automation. International Journal of Engineering Science and Advanced Technology, 20(4), 21–26. https://doi.org/10.64771/ijesat.2020.v20.i04.pp3846
60. Ahuja, R. (2022). Application of static application security testing (SAST) in CI/CD pipelines for early detection of insecure coding practices through syntax tree analysis and custom rule sets. International Journal on Recent and Innovation Trends in Computing and Communication, 10(2), 80–87. https://www.ijritcc.org/index.php/ijritcc/article/view/11922

HOW TO CITE THIS ARTICLE: Yusupov, S., & Jain, K. (2025). A Hybrid Speech–Text Foundation Model for Noise-Robust Multilingual Automatic Speech Recognition. *Journal of Integrated Science, Technology and Management*, 1(2), 21-29.