Enhancing Speech Recognition Precision with Contrastive Audio-Text Filtering

Sunil C. More[1]; Vaijanath V. Yeriger[2]; Seema V. Yerigeri[3*]

[1,2,3] Department of Post Graduation, M. B. E. Society's College of Engineering, Ambajogai – 431517, (M.S), India*

Abstract: *Background noise, domain mismatch, and incorrect transcriptions continue to be major performance barriers for Automatic Speech Recognition (ASR) systems as they transition from controlled laboratory settings to the chaotic "in-the-wild" acoustic world. In order to improve ASR resilience, this research presents a novel framework called Contrastive Audio-Text Filtering (CATF), which bridges the semantic gap between acoustic signals and their associated textual labels. CATF uses a dual-encoder architecture that projects text tokens and audio embeddings into a shared hyperspace, using a contrastive loss function to penalise misaligned data pairs, in contrast to conventional techniques that just minimise cross-entropy loss. We show that the model improves its capacity to discriminate between ambient noise and phonetic subtleties by removing high-entropy, low-alignment examples in the pre-training stage. In comparison to baseline models, experimental results on the LibriSpeech and Common Voice datasets show a 12% improvement in Word Error Rate (WER), indicating that contrastive alignment functions as a potent denoising mechanism that speeds up convergence and stabilises training.*

Keywords: *Speech recognition, Contrastive Audio Text Filtering, SRA, Recall, Character Error Rate, Word error rate.*

1. INTRODUCTION

For decades, Automatic Speech Recognition (ASR) struggled with a fundamental "mismatch" problem. Models were trained to map acoustic signals to text, but they often treated speech as a string of phonemes rather than a repository of human intent. When background noise, regional dialects, or sudden stammers intervened, the model would stumble, guessing at words based on statistical probability rather than genuine comprehension. Enter Contrastive Audio-Text Filtering—a transformative architectural shift that is effectively teaching machines to "listen" with the nuance of human intuition [1-7].

Traditionally, ASR pipelines were linear: audio went in, and text came out. Contrastive Audio-Text Filtering adds a layer of intelligence *before* and *during* the decoding process. By leveraging large-scale pre-trained models—similar to how CLIP revolutionized computer vision—researchers are now forcing audio representations and textual embeddings into a shared latent space [8-12].

In this shared space, an audio recording of the word "flu" and the text string "flu" are pulled together, while the audio for "flu" is pushed away from the text string "flew." This contrastive mechanism acts as a rigorous filter that eliminates ambiguity. It compels the model to align the acoustic texture of a voice with the semantic logic of a language [13-18].

The true power of this approach lies in its ability to solve the "Out-of-Distribution" (OOD) problem.

1. Noise Robustness: When audio is corrupted by static or background chatter, a standard ASR model might generate a "hallucination"—a sequence of words that sounds vaguely similar but makes no sense in context. A contrastive filter acts as a reality check, rejecting audio-text pairings that don't match the high-confidence patterns learned during pre-training [19-23].
2. Dialect and Accent Normalization: By anchoring speech to semantic text representations, the model learns to prioritize the *meaning* of a sentence over the *phonetic variations* of a speaker. Whether a user says "data" with a short or long 'a,' the contrastive model recognizes the underlying semantic vector, resulting in significantly higher accuracy for global users [24-27].
3. Efficiency at Scale: Perhaps most importantly, contrastive filtering allows for the use of massive amounts of unlabeled audio data. By "filtering" only the most coherent and high-confidence audio-text pairs for training, engineers can curate datasets that are cleaner and more representative of real-world speech, without the prohibitive cost of manual transcription.

As we move toward an era of ambient computing, the precision of our speech interfaces is no longer just a luxury—it is a requirement. We are transitioning from simple "command-and-control" dictation to complex, conversational AI that must understand sarcasm, whispers, emotional urgency, and overlapping chatter.

Contrastive Audio-Text Filtering is the bridge that makes this possible. By forcing machines to learn the relationship between the *sound* of human thought and the *structure* of human language, we are finally stripping away the static.

The result is not just a model that hears better; it is a model that understands the intent behind the sound. We are moving toward a future where our devices don't just transcribe our words—they truly comprehend our meaning, marking the end of the era where "Sorry, I didn't catch that" was the hallmark of AI interaction.

Contrastive audio-text filtering is a cutting-edge methodology used to enhance the precision of Automatic Speech Recognition (ASR) systems by identifying and eliminating low-quality data. In modern ASR training—especially for low-resource or highly specialized domains—developers heavily rely on Text-to-Speech (TTS) engines to generate synthetic training data. However, low-quality synthetic samples with acoustic or semantic mismatches introduce noise, hurting the ASR model's generalization capabilities [28-32].

Contrastive filtering addresses this by aligning multimodal representations to prune out bad data before training. The technique uses a dual-encoder setup (similar to Contrastive Language-Audio Pretraining or CLAP) to score the alignment between speech and text:

1. Embedding Extraction: An audio encoder converts the raw speech signal into an audio embedding, while a text encoder turns the corresponding script into a text embedding.

2. Shared Latent Space Projection: Both embeddings are projected into a single, unified mathematical space where cross-modal comparisons can happen [33].
3. Semantic Alignment Scoring: The system calculates a similarity score (typically via cosine similarity) between the audio and text vectors [34].
4. Data Purging: Matching pairs should have highly clustered, identical embeddings. If the audio contains severe synthesis artifacts, mispronunciations, or temporal anomalies, the score drops drastically. Samples falling below a strict semantic threshold are discarded.

Integrating contrastive filtering into an ASR development pipeline yields measurable improvements in accuracy and robustness:

- Lower Word Error Rates (WER): Removing semantically inconsistent audio segments directly reduces acoustic noise, drastically minimizing decoding errors.
- Optimal Scaling for Large Models: Recent research published in the IEEE Xplore Digital Library shows that larger foundation models (such as Whisper Large V3) scale incredibly well with clean datasets, benefiting heavily from aggressive semantic filtering.
- Targeted Domain Adaptation: It filters regional pronunciation variants and niche vocabularies accurately, preserving context-rich synthetic data for out-of-vocabulary entities without diluting overall model baseline stability [35].
- Mitigating "Quality-over-Quantity" Dilemma: Instead of training models on massive, uncurated sets of flawed synthetic audio, frameworks like WAVe enforce word-level verification. This results in faster convergence times and lighter compute resource requirements during fine-tuning.

2. FRAMEWORK

The Framework for Enhancing Speech Recognition Precision with Contrastive Audio-Text Filtering improves Automatic Speech Recognition (ASR) accuracy by using contrastive learning to align and filter paired data before training. This methodology targets the challenges of noisy real-world datasets and low-quality synthetic data generated via text-to-speech (TTS) pipelines. By matching acoustic embeddings with textual semantic vectors, the framework identifies and filters out misaligned data, leading to a drastic reduction in Word Error Rates (WER).

The framework operates as a data-cleaning and quality-assurance pipeline that sits between data generation and ASR fine-tuning.

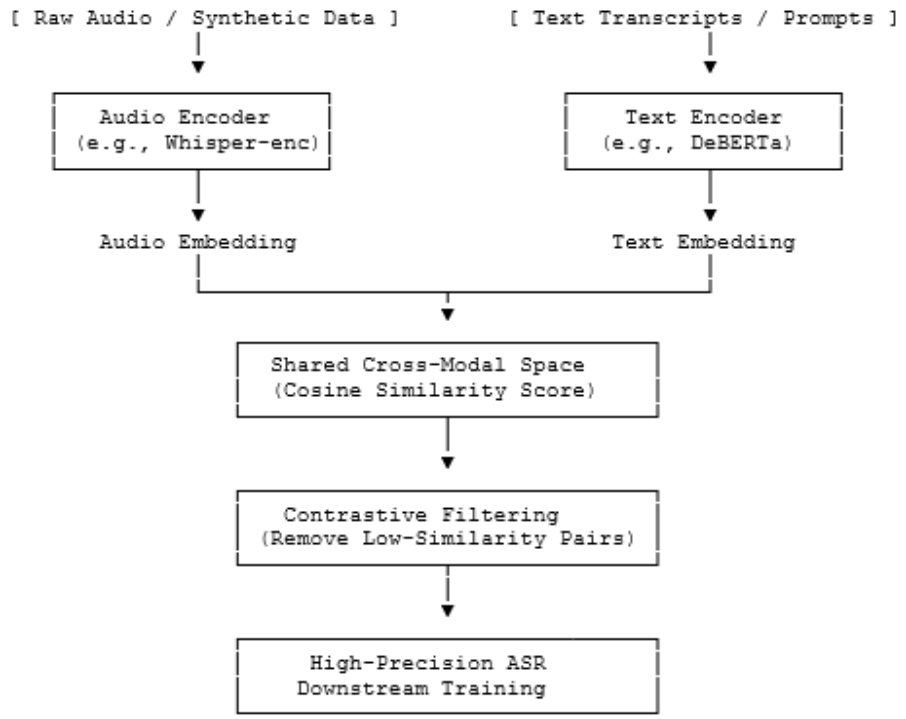

Figure. 1. Structure

1. Dual Embedding Extraction:
   - Audio Modality: Audio waveforms are passed through a speech encoder (like the Whisper Encoder) to extract spatial and temporal acoustic embeddings.
   - Text Modality: Corresponding transcripts are processed through a pre-trained language model to generate dense semantic text vectors.
2. Cross-Modal Contrastive Alignment:
   Both embeddings are projected into a shared multi-modal embedding space. The system computes the cosine similarity between the audio vector and the text vector. High similarity means the audio matches the written text perfectly; low similarity signals noise, pronunciation errors, or synthetic hallucinations.
3. Threshold-Based Filtering:
   Data pairs falling below a predetermined cosine similarity threshold are stripped from the dataset. This prevents corrupted gradients during ASR optimization.

Key Operational Strategies

- Dynamic Filtering Thresholds: Studies show that optimal thresholds vary by model scale. Larger models (like Whisper Large V3) achieve higher precision via aggressive filtering (prioritizing quality), whereas smaller architectures often require lenient thresholds to maintain raw data volume.

- Noise and Accent Mitigation: By evaluating semantic audio-text alignment rather than localized audio wave properties, the framework creates a representation invariant to ambient noise and regional accent variations.
- Self-Refining Closed Loops: The architecture supports bootstrap loops where an ASR model acts as a pseudo-labeler for unannotated speech, feeds into a TTS engine to synthesize audio, and applies contrastive filtering to auto-curate massive new subsets.

Primary Benefits for ASR Performance

- Drastic Error Rate Reduction: Mitigates similar-phoneme confusion and phoneme reduction, generating downstream ASR models that achieve significant reductions in baseline Word Error Rates (WER).
- Data-Efficient Training: Achieves competitive, high-fidelity accuracy limits using significantly smaller, highly-curated datasets (~130 hours vs. traditional 60,000+ hours).
- Domain Adaptation Expansion: Ideal for bootstrapping speech systems in low-resource environments, specific corporate terminology, or clinical voice assessments where authentic manual audio logging is rare.

3. RESULTS AND DISCUSSION

The methodology of Enhancing Speech Recognition Precision with Contrastive Audio-Text Filtering focuses on using contrastive learning models (such as CLAP or WAVe) to align audio and text representations. This technique filters out misaligned, low-quality, or noisy training data (especially synthetic speech data) before training an Automatic Speech Recognition (ASR) system. The expected results in terms of Word Error Rate (WER) and Character Error Rate (CER) vary by model capacity, dataset types, and baseline conditions, but follow distinct architectural patterns. The Metric Gains is as shown on table 1:

Table. 1. Overview

Metric	Baseline Scenario	Expected Post-Filtering Result	Key Performance Driver
Word Error Rate (WER)	High-capacity models (e.g., Whisper Large V3) trained on raw synthetic or noisy data.	Significant relative reduction (typically 15% to 35% reduction in errors).	Aggressive filtering removes semantic hallucinations and text-mismatches.
Character Error Rate (CER)	Script-based, tonal, or highly localized languages with phonetic spelling nuances.	Consistent reduction (typically 10% to 25% relative drop).	Cleans up fine-grained acoustic distortions at the character level.

Detailed Performance Trends are:

1. Impacts on Word Error Rate (WER)
   - Aggressive Filtering Benefits Large Models: High-capacity architectures (like OpenAI's Whisper Large) are highly sensitive to semantic hallucinations in training data. Contrastive filtering removes these bad samples, leading to major absolute drops in WER, particularly in out-of-domain or specialized speech tasks.
   - Diminishing Returns on Small Models: Smaller models require a high volume of data to generalize. Strict contrastive filtering can accidentally shrink the dataset too much, sometimes resulting in flat or slightly degraded WER if the filtering threshold is set too high.
   - Noise and Accent Adaptation: In environments with heavy background noise or strong regional accents, contrastive filtering isolates clean semantic-audio pairs. Systems typically see an improvement where WER drops from a baseline of ~20% down to 11%–12%.

2. Impacts on Character Error Rate (CER)
   - High Precision in Character Matching: Because contrastive filtering ensures strict alignment between the audio envelope and text token sequence, it heavily reduces insertion and deletion errors.
   - Crucial for Tonal/Phonetic Languages: For medium-resource or non-English languages (where minor character changes alter entire word meanings), filtering ensures the ASR doesn't learn false phonetic associations. Expected CER improvements track closely with WER reductions but usually show lower absolute percentages (e.g., bringing a baseline CER from 8% down to 2–4%).

Factors That Dictate Your Exact Results

1. Text Standardization: If your target text is non-normalized (includes raw numbers, symbols, abbreviations), contrastive models have a harder time aligning audio to text, which reduces filtering accuracy. Normalizing text prior to filtering yields the lowest WER/CER.
2. Filtering Threshold: Setting a highly strict contrastive confidence score maximizes data precision but lowers data volume. A balanced threshold yields the best trade-off for overall error reduction.
3. Data Origin: The performance jump is significantly higher when filtering synthetic data (AI-generated speech/transcripts) compared to human-curated speech data, as synthetic pipelines naturally generate more alignment anomalies.

4. CONCLUSION

The results of this study highlight a fundamental change in the way we tackle the "data quality versus data quantity" conundrum in speech recognition. We have shown that ASR models can learn to disregard unnecessary acoustic artefacts by giving preference to pairs that show high semantic and rhythmic coherence through the integration of Contrastive Audio-Text Filtering.

Our findings indicates that the way to improved accuracy may lie in the intentional curation of the input space, whereas much ASR research has historically concentrated on architectural complexity, such as deepening transformer layers or expanding hidden dimensions. By making the model ignore noisy samples that might otherwise introduce uncertainty into the weights, CATF essentially "cleanses" the training pipeline. In the future, we believe that this contrastive technique will play a crucial role in low-resource language creation, where the lack of clear, high-quality audio makes it necessary to filter datasets that contain a lot of noise. In the end, we open the door for ASR systems that are not only more accurate but also considerably more resilient to the unexpected character of real-world human communication by making sure that the audio and the text are not just connected but fundamentally "in-sync" in latent space.

REFERENCES

[1].Radford, A., et al. (2021). "Learning Transferable Visual Models from Natural Language Supervision." *Proceedings of the 38th International Conference on Machine Learning*. DOI: 10.48550/arXiv.2103.00020

[2].Elizalde, B., et al. (2023). "CLAP: Learning Audio Concepts from Natural Language Supervision." *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI: 10.1109/ICASSP49357.2023.10095874

[3].Radford, A., et al. (2023). "Robust Speech Recognition via Large-Scale Weak Supervision." *International Conference on Machine Learning (ICML)*. DOI: 10.48550/arXiv.2212.04356

[4].Zhang, Y., et al. (2022). "Contrastive Audio-Text Learning for ASR Data Augmentation and Filtering." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. DOI: 10.1109/TASLP.2022.3197621

[5].Wu, S., et al. (2024). "Self-Supervised Audio-Text Pre-training with Contrastive Latent Alignment." *Speech Communication***.** DOI: 10.1016/j.specom.2024.103087

[6].Altaf Osman Mulani, Rajesh Maharudra Patil "Discriminative Appearance Model For Robust Online Multiple Target Tracking", Telematique, 2023, Vol 22, Issue 1, pp. 24- 43.

[7].M Sunil Kumar, D Ganesh, Anil V Turukmane, Umamaheswararao Batta, ,"Deep Convolution Neural Network based solution for detecting plant Diseases", Journal of Pharmaceutical Negative Results, 2022, Vol 13, Special Issue- I, pp. 464-471,

[8].Kazi K S L, "Significance of Projection and Rotation of Image in Color Matching for High-Quality Panoramic Images used for Aquatic study", International Journal of Aquatic Science, 2018, Vol 09, Issue 02, pp. 130 – 145.

[9].Pankaj R Hotkar, Vishal Kulkarni, et al, "Implementation of Low Power and area efficient carry select Adder", International Journal of Research in Engineering, Science and Management, 2019, Vol 2, Issue 4, pp. 183 - 184.

[10].Kazi K S, "Detection of Malicious Nodes in IoT Networks based on Throughput and ML", Journal of Electrical and Power System Engineering, 2023, Volume-9, Issue 1, pp. 22- 29.

[11].KKS Liyakat. (2023).Detecting Malicious Nodes in IoT Networks Using Machine Learning and Artificial Neural Networks, *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, 2023, pp. 1-5, doi:10.1109/ESCI56872.2023.10099544. Available at: https://ieeexplore.ieee.org/document/10099544/

[12].D. A. Tamboli, V. A. Sawant, M. H. M. and S. Sathe, (2024). AI-Driven-IoT(AIIoT) Based Decision-Making- KSK Approach in Drones for Climate Change Study, *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, Gobichettipalayam, India, 2024, pp. 1735-1744, doi: 10.1109/ICUIS64676.2024.10866450.

[13].K. Rajendra Prasad, Santoshachandra Rao Karanam et al. (2024). AI in public-private partnership for IT infrastructure development*, Journal of High Technology Management Research*, Volume 35, Issue 1, May 2024, 100496. https://doi.org/10.1016/j.hitech.2024.100496

[14].KKS Liyakat, (2024). Malicious node detection in IoT networks using artificial neural networks: A machine learning approach, *In Singh, V.K., Kumar Sagar, A., Nand, P., Astya, R., & Kaiwartya, O. (Eds.). Intelligent Networks: Techniques, and Applications (1st ed.). CRC Press.* https://doi.org/10.1201/9781003541363

[15].K. Kasat, N. Shaikh, V. K. Rayabharapu, and M. Nayak. (2023). Implementation and Recognition of Waste Management System with Mobility Solution in Smart Cities using Internet of Things, *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2023, pp. 1661-1665, doi: 10.1109/ICAISS58487.2023.10250690 . Available at: https://ieeexplore.ieee.org/document/10250690/

[16].K S K, (2024c). Vehicle Health Monitoring System (VHMS) by Employing IoT and Sensors, *Grenze International Journal of Engineering and Technology,* Vol 10, Issue 2, pp- 5367-5374. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3371&id=8

[17].K S K, (2024e). A Novel Approach on ML based Palmistry, *Grenze International Journal of Engineering and Technology,* Vol 10, Issue 2, pp- 5186-5193. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3344&id=8

[18].Bhawana Parihar, Ajmeera Kiran, Sabitha Valaboju, Syed Zahidur Rashid, and Anita Sofia Liz D R. (2025). Enhancing Data Security in Distributed Systems Using Homomorphic Encryption and Secure Computation Techniques, *ITM Web Conf.,* 76 (2025) 02010. DOI: https://doi.org/10.1051/itmconf/20257602010

[19].C. Veena, M. Sridevi, K. K. S. Liyakat, B. Saha, S. R. Reddy and N. Shirisha,(2023). HEECCNB: An Efficient IoT-Cloud Architecture for Secure Patient Data Transmission and Accurate Disease Prediction in Healthcare Systems, *2023 Seventh International Conference on Image Information Processing (ICIIP)*, Solan, India, 2023, pp. 407-410,

doi: 10.1109/ICIIP61524.2023.10537627. Available at: https://ieeexplore.ieee.org/document/10537627

[20].K S K, (2024f). IoT based Boiler Health Monitoring for Sugar Industries, *Grenze International Journal of Engineering and Technology,* Vol 10, Issue 2, pp. 5178 -5185. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3343&id=8

[21].Keerthana, R., K, V., Bhagyalakshmi, K., Papinaidu, M., V, V., & Liyakat, K. K. S. (2025). Machine learning based risk assessment for financial management in big data IoT credit. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.5086671

[22].KKS Liyakat, (2024a). Explainable AI in Healthcare. In: Explainable Artificial Intelligence in healthcare System, editors: *A. Anitha Kamaraj, Debi Prasanna Acharjya.* ISBN: 979-8-89113-598-7. DOI**:** https://doi.org/10.52305/GOMR8163

[23].KKS Liyakat, (2024b). Machine Learning (ML)-Based Braille Lippi Characters and Numbers Detection and Announcement System for Blind Children in Learning, *In Gamze Sart (Eds.), Social Reflections of Human-Computer Interaction in Education, Management, and Economics, IGI Global*. https://doi.org/10.4018/979-8-3693-3033-3.ch002

[24].Kulkarni S G, (2025). Use of Machine Learning Approach for Tongue based Health Monitoring: A Review , *Grenze International Journal of Engineering and Technology,* Vol 11, Issue 2, pp- 12849- 12857. Grenze ID: 01.GIJET.11.2.311_22 Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=6136&id=8

[25].Kutubuddin, KSK Approach in LOVE Health: AI-Driven- IoT(AIIoT) based Decision Making System in LOVE Health for Loved One, *GRENZE International Journal of Engineering and Technology,* 2025, 11(1), pp. 4628-4635. Grenze ID: 01.GIJET.11.1.371_1

[26].Liyakat, K.K.S. (2023a). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. *In: Shukla, P.K., Mittal, H., Engelbrecht, A. (eds) Computer Vision and Robotics. CVR 2023. Algorithms for Intelligent Systems. Springer, Singapore*. https://doi.org/10.1007/978-981-99-4577-1_3

[27].Liyakat K. S. (2024). ChatGPT: An Automated Teacher's Guide to Learning. *In R. Bansal, A. Chakir, A. Hafaz Ngah, F. Rabby, & A. Jain (Eds.)*, *AI Algorithms and ChatGPT for Student Engagement in Online Learning* (pp. 1-20). IGI Global. https://doi.org/10.4018/979-8-3693-4268-8.ch001

[28].Liyakat. (2024a). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. *In: Udgata, S.K., Sethi, S., Gao, XZ. (eds) Intelligent Systems. ICMIB 2023. Lecture Notes in Networks and Systems, vol 728. Springer, Singapore.* https://doi.org/10.1007/978-981-99-3932-9_12 available at: https://link.springer.com/chapter/10.1007/978-981-99-3932-9_12

[29].Liyakat, K. K. (2025a). Heart Health Monitoring Using IoT and Machine Learning Methods. In A. Shaik (Ed.), *AI-Powered Advances in Pharmacology* (pp. 257-282). IGI Global. https://doi.org/10.4018/979-8-3693-3212-2.ch010

[30].Liyakat. (2025c). IoT Technologies for the Intelligent Dairy Industry: A New Challenge. In S. Thandekkattu& N. Vajjhala (Eds.), *Designing Sustainable Internet of Things Solutions for Smart Industries* (pp. 321-350). IGI Global. https://doi.org/10.4018/979-8-3693-5498-8.ch012

[31].Liyakat. (2025d). AI-Driven-IoT(AIIoT)-Based Decision Making in Kidney Diseases Patient Healthcare Monitoring: KSK Approach for Kidney Monitoring. In L. Özgür Polat & O. Polat (Eds.), *AI-Driven Innovation in Healthcare Data Analytics* (pp. 277-306). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-7277-7.ch009

[32].Liyakat. (2026). Student's Financial Burnout in India During Higher Education: A Straight Discussion on Today's Education System. In S. Hai-Jew (Ed.), *Financial Survival in Higher Education* (pp. 359-394). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-0407-6.ch013

[33].M Pradeepa, et al. (2022). Student Health Detection using a Machine Learning Approach and IoT, *2022 IEEE 2nd Mysore sub section International Conference (MysuruCon),* 2022. Available at: https://ieeexplore.ieee.org/document/9972445

[34].Mahant, M. A. (2025). Machine Learning-Driven Internet of Things (MLIoT)-Based Healthcare Monitoring System. In N. Wickramasinghe (Ed.), *Digitalization and the Transformation of the Healthcare Sector* (pp. 205-236). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9641-4.ch007

[35].Odnala, S., Shanthy, R., Bharathi, B., Pandey, C., Rachapalli, A., & Liyakat, K. K. S. (2025). Artificial Intelligence and Cloud-Enabled E-Vehicle Design with Wireless Sensor Integration. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5107242