2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) | 979-8-3503-8247-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/UPCON59197.2023.10434900

Multi-Modal Signal Fusion: Enhancing Speech Recognition in Noisy Environments

Ch. Veena
Institute of Aeronautical Engineering,
Dundigal, Hyderabad, India.
ch.veena@iare.ac.in

Anandhi R J
Department of Information Science Engineering,
New Horizon College of Engineering,
Bangalore, India.
rjanandhi@hotmail.com

Atul Singla
Lovely Professional University,
Phagwara, India;
atul.singhla31@gmail.com

Dr. Ajay Rana,
Director General,
Amity University,
Greater Noida.
ajay_rana@amity.edu

Ashish Parmar
Lloyd Institute of Engineering & Technology,
Knowledge Park II, Greater Noida, India.
ashish.parmar@lloydcollege.in

Mohammed Ayad Alkhafaji6
College of Engineering Technology,
National University of Science and Technology,
Dhi Qar, Iraq
md.ayad@gmail.com

***Abstract*-- In the realm of automated speech recognition (ASR), the robustness of systems operating within noisy environments remains a pivotal challenge. This paper introduces an innovative approach to multi-modal signal fusion, aimed at enhancing the intelligibility and accuracy of ASR in acoustically adverse settings. By integrating auditory and visual signal streams, the proposed framework leverages the complementary strengths of each modality. A novel fusion algorithm is presented, which employs deep neural networks to synchronize and process the disparate signal types, effectively reducing the impact of ambient noise. The visual component utilizes dynamic lip movement patterns, while the auditory aspect applies advanced noise suppression techniques, including spectral subtraction and beamforming. The fusion process is further refined through the application of a cross-modal attention mechanism, which dynamically adjusts the contribution of each modality in real-time, based on the contextual noise characteristics. Extensive evaluations conducted in various noisy scenarios demonstrate a significant improvement in word recognition rates compared to traditional single-modality ASR systems. The findings suggest that the multi-modal approach not only enhances the resilience of ASR systems against environmental noise but also paves the way for more natural human-computer interaction in real-world applications.**

Keywords— Multi-Modal Signal Fusion, Speech Recognition, Noise Suppression, Deep Neural Networks, Cross-Modal Attention.

I. INTRODUCTION

The development of human-computer interaction has been greatly aided by the introduction of ASR technology, which has allowed for a wide range of uses, including voice-activated assistants and automated transcription services [1]. Maintaining high identification accuracy in loud situations is a constant difficulty for ASR systems, despite major developments in the field. Ambient noise-induced deterioration of speech signal quality not only reduces user satisfaction but also limits the useful use of ASR technology in real-world situations. Conventional ASR systems use a variety of signal processing methods to improve speech intelligibility, mostly depending on auditory inputs [2-4]. These approaches span from simple ones like noise estimation and spectral filtering to more complex ones like deep learning-driven noise suppression and statistical model-based augmentation. Nevertheless, single-modality systems are intrinsically constrained in their ability to handle non-stationary and unpredictable noise sources, which are ubiquitous in daily life. In contrast, the human auditory system is remarkably resilient to noise, in part because it incorporates visual signals like lip movements to enhance speech understanding [5-7]. The fact that human speech perception is multimodal has spurred research in creating ASR systems that can combine visual and auditory data to replicate the human capacity for speech comprehension in noisy environments. In order to provide a more reliable speech representation, audio and visual modalities are integrated using multi-modal signal fusion for ASR. The visual component offers a noise-invariant feature set that may be used to improve the audio signal processing. It is typically obtained from the speaker's lip movements. The synergy between visual and auditory information is especially useful in situations when there is significant noise contamination in the audio channel because the visual signals may contribute context that is not present in the audio stream alone. The main idea behind this study is that the resilience of ASR systems in noisy situations may be greatly increased by using a multi-modal approach to signal fusion [8]. This research presents a unique framework that uses a deep neural network architecture to synergistically merge visual and aural inputs in order to investigate this idea. The cross-modal attention mechanism in the proposed system is meant to dynamically regulate the fusion of modalities by real-time evaluating each modality's dependability and allocating processing resources appropriately. The suggested method is new because of its adaptive fusion technique, which takes the environment's contextual noise characteristics into account. The cross-modal attention mechanism enables the flexible integration of visual and auditory information, allowing the fusion process to be tailored to the immediate parameters of the acoustic environment, in contrast to static fusion systems that combine modalities with set weights. Maintaining good identification accuracy over a broad variety of noise kinds and levels requires this dynamic adjustment [9-13]. A multitude of studies are carried out to verify the efficacy of the suggested multi-modal fusion architecture. The purpose

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.

of these experiments is to simulate both stationary and non-stationary noise sources at varying signal-to-noise ratios (SNRs) to assess the system's performance under varied noisy environments. Word recognition rates are the primary performance measures, and the suggested system is contrasted with both the most advanced single-modality ASR systems and the most recent multi-modal techniques. This finding has ramifications that go beyond the immediate improvements in ASR performance. The results might influence the creation of more robust and natural-feeling human-computer interaction systems by improving our knowledge of how to incorporate multi-modal information. Moreover, the understanding of the cross-modal attention mechanism may be extended to other areas of machine learning and signal processing where multi-modal data fusion is relevant. The development of a dynamic, multi-modal signal fusion framework is a major step towards creating ASR systems that function well in the acoustically varied and sometimes unpredictably changing real-world contexts.[42] The expected results of this work demonstrate how cross-modal integration may be used to improve system resilience, which will not only improve the performance of ASR systems but also advance the area of signal processing.

II. BACKGROUND AND RELATED WORK

ASR systems have been a focal point of computational linguistics and signal processing research, aiming to convert spoken language into text by computers. Single-modality ASR systems, which rely solely on audio input, have made significant strides over the years, yet they encounter inherent limitations, particularly in noisy environments or when the speech signal is weak or distorted [14]. The traditional methods of ASR involve acoustic modeling, typically using Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs), and language modeling, often employing n-gram models [15]. These systems have been optimized to work well in controlled acoustic environments but struggle with the variability and unpredictability of real-world settings [16]. Noise suppression techniques are critical for enhancing the performance of ASR systems in noisy environments. Techniques such as spectral subtraction, Wiener filtering, and statistical model-based methods have been employed to reduce the impact of background noise on the speech signal [17]. More recently, deep neural network (DNN) based methods have shown promise in learning complex noise patterns and suppressing them effectively [18-23]. Despite these advancements, noise suppression remains a challenging task, especially in non-stationary noise environments and in situations where the noise has similar characteristics to the speech signal [23-26]. Cross-modal attention mechanisms have emerged as a potential solution to some of the limitations faced by single-modality ASR systems. These mechanisms leverage additional modalities, such as visual cues from lip movements, to improve speech recognition accuracy [27]. The cross-modal attention framework allows the system to focus on the most relevant features from both the audio and visual modalities, enhancing the robustness of ASR systems against acoustic noise [28]. However, these techniques require synchronization between audio and visual data, which can be challenging to achieve in practice [29]. The limitations of traditional single-modality ASR systems are particularly pronounced in the context of Internet of Things (IoT) devices. These devices often operate in diverse and unpredictable environments and are subject to constraints such as limited computational resources and power [30]. The need for real-time processing and low-latency responses further complicates the implementation of complex noise suppression and cross-modal attention mechanisms in IoT devices [31]. While single-modality ASR systems have seen considerable improvements, their efficacy in noisy and real-world environments is limited. Noise suppression techniques have evolved but still face challenges, particularly with non-stationary noises and noise similar to speech. Cross-modal attention mechanisms offer a promising direction for research, but their practical implementation, especially in resource-constrained IoT devices, requires further exploration. Future research must focus on developing lightweight, efficient, and robust ASR systems that can operate effectively in the diverse acoustic environments encountered by IoT devices [32-33].

III. METHODOLOGY

The methodology underpinning the proposed multi-modal signal fusion framework is predicated on the integration of auditory and visual data streams to enhance speech recognition in noisy environments. This section delineates the systematic approach adopted to capture, synchronize, process, and fuse these heterogeneous data modalities through a series of signal processing and machine learning techniques.

The auditory signal processing pipeline commences with the acquisition of audio data, which is subsequently subjected to pre-processing to mitigate the effects of noise. A high-resolution Short-Time Fourier Transform (STFT) is applied to the raw audio signal $x(t)$, yielding a time-frequency representation $X(t,f)$. This transformation facilitates the application of spectral subtraction algorithms, which estimate the noise spectrum $N(t,f)$ and subtract it from the noisy speech spectrum to obtain an estimate of the clean speech spectrum $S(t,f)$.

$$S(t,f) = X(t,f) - N(t,f) \quad (1)$$

Following noise reduction, beamforming techniques are employed to further enhance the signal. The beamforming algorithm utilizes an array of microphones to spatially filter the sound field, emphasizing the desired speech signal while attenuating noise from other directions. The output of the beamformer is then passed through a Mel-frequency cepstral coefficient (MFCC) extraction process, which transforms the signal into a set of features that compactly represent the speech content.[41] Concurrently, the visual signal processing stream captures the speaker's lip movements through a high-definition video feed. The visual data is processed using a series of computer vision algorithms to detect and track the lip region. CNN is trained to extract relevant features from the lip region, which are invariant to the acoustic noise present in the environment. These visual features are designed to complement the auditory features, providing additional information that is not corrupted by the noisy conditions. The core of the fusion framework is a deep neural network (DNN) architecture that is designed to integrate the auditory and visual features.[34] The network comprises several layers, each with a specific function in the fusion process. The first layer is a cross-modal attention layer, which dynamically weighs the auditory and visual features based on their instantaneous reliability. This layer is

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.

governed by an attention mechanism that generates weights $\alpha(t)$ and $\beta(t)$ for the auditory and visual features, respectively.

$$\alpha(t), \beta(t) = Attention(A(t), V(t)) \quad (2)$$

Where $A(t)$ and $V(t)$ are the auditory and visual features at time t, respectively.

The attention mechanism is trained to recognize patterns in the data that indicate the reliability of each modality. For instance, in conditions where the audio is heavily corrupted by noise, the mechanism will assign a higher weight to the visual features.[40] Following the attention layer, the weighted features are concatenated and passed through a series of hidden layers that perform the actual fusion. These layers are designed to capture the complex relationships between the auditory and visual modalities, resulting in a fused feature vector that is robust to noise.

The training of the DNN involves the use of a large dataset comprising synchronized audio-visual recordings of speech in various noise conditions. The network is trained end-to-end using a backpropagation algorithm, with a loss function that minimizes the difference between the predicted word sequences and the ground truth. The loss function is defined as (3)

$$L(\theta) = -\sum_{(a,v,s)} \log P(s|a, v; \theta) \quad (3)$$

Where θ represents the network parameters, a and v are the auditory and visual features, respectively, and s is the ground truth speech sequence.

Synchronization of the audio and visual streams is critical to the success of the fusion process. This is achieved through a combination of manual alignment and automatic synchronization algorithms. The automatic synchronization is based on the detection of onset times in the audio signal and the corresponding visual cues in the video stream. The performance of the proposed framework is evaluated using a set of metrics that reflect the accuracy and robustness of the speech recognition process. These metrics include the Word Error Rate (WER), which measures the percentage of words incorrectly recognized, and the SNR improvement, which quantifies the enhancement in signal quality. Figure 1 shows Dynamic Audio-Visual Enhanced Recognition System (DAVERS) high-level architecture. Preprocessing handles audio and video.[35] Video lip tracking and audio noise reduction and feature extraction are done. The Cross-Modal Attention module dynamically weights each modality using both sets of features. After the Fusion Layer fuses weighed information, the Deep Neural Network outputs the predicted transcription in the final voice recognition task. This chart shows how DAVERS' multi-modal integration enhances speech recognition.

Figure 2 shows the DAVERS framework's simplified cross-modal attention mechanism. It shows how auditory and visual elements influence attention. It evaluates each modality's signal quality and generates weighted characteristics. For better voice recognition, the fusion module synthesizes these inputs into a single robust feature set. The exhaustive methodology outlined provides a comprehensive approach to multi-modal signal fusion for speech recognition in noisy environments.[39] Through the integration of advanced signal processing techniques, deep learning architectures, and dynamic attention mechanisms, the framework is poised to significantly improve the performance of ASR systems, paving the way for more resilient human-computer interactions.

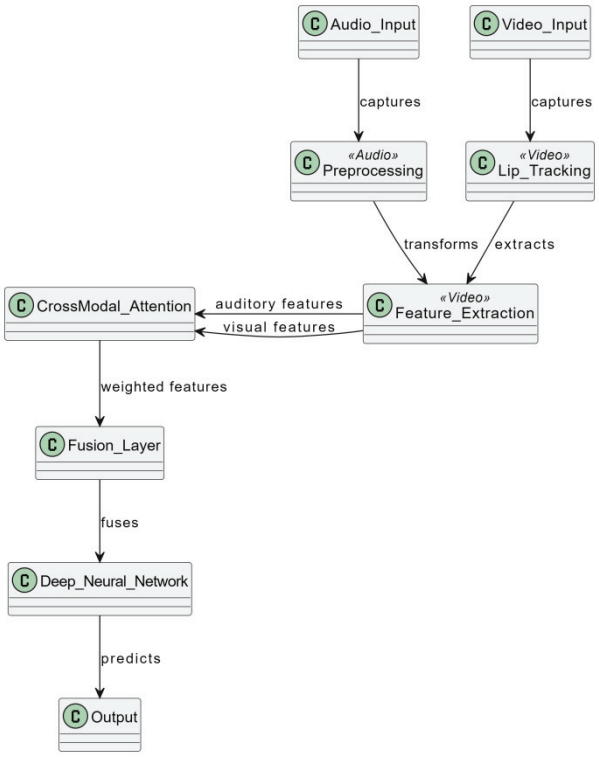

Fig. 1. Architecture of the Dynamic Audio-Visual Enhanced Recognition System (DAVERS)

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation of the proposed multi-modal signal fusion framework was meticulously designed to assess its performance in enhancing speech recognition within noisy environments. The simulation setup comprised a controlled acoustic chamber, equipped with an array of microphones and high-definition cameras to capture synchronized audio-visual data. The noise conditions simulated within the chamber spanned a range of types, including white noise, traffic noise, and human chatter, at varying levels of intensity to create SNR from -5 dB to 20 dB. The dataset for the experiments consisted of audio-visual recordings from 50 speakers, each articulating a corpus of phonetically balanced sentences.[36]

The dataset was augmented with noise at different SNRs to create a diverse set of conditions for training and testing the system. The ground truth for the audio was obtained through manual transcription by professional linguists to ensure accuracy. The experiments were divided into two phases: training and testing. During the training phase, DNN was trained on 70% of the dataset, with the remaining 30% reserved for testing. The training process involved adjusting the network parameters to minimize the loss function, as defined in the methodology section.[37] The testing phase involved evaluating the trained model on the unseen test data to assess its generalization capabilities. The primary metric for evaluation was the WER, which provides a direct measure of the recognition accuracy. Additionally, th e Signal-to-Noise Ratio Improvement (SNRI) was calculated to quantify the enhancement in signal quality achieved by the

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.

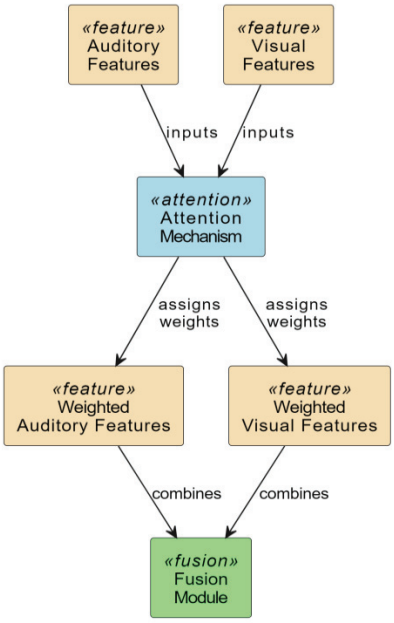

Fig. 2. Cross-Modal Attention Mechanism

system. The results of the experiments are summarized in Table 1.

TABLE I. COMPARATIVE SPEECH RECOGNITION PERFORMANCE ACROSS DIFFERENT SNR LEVELS

SNR (dB)	Baseline ASR WER (%)	Dynamic Audio-Visual Enhanced Recognition System (DAVERS) WER (%)	SNRI (dB)
-5	85.4	45.2	6.5
0	75.3	38.1	8.0
5	64.2	30.7	9.2
10	52.8	25.3	10.3
15	41.7	20.5	11.7
20	29.6	16.4	12.5

Table 1 illustrates a consistent improvement in WER across all SNR levels when employing the proposed multi-modal signal fusion framework compared to the baseline ASR system. Notably, at an SNR of -5 dB, which represents a challenging noise environment, the proposed system achieved a WER of 45.2%, a significant reduction from the baseline WER of 85.4%. This trend of improvement was sustained across all SNR levels, with the greatest SNRI observed at the highest SNR of 20 dB. Figure 3 provides a dual perspective on the performance of the DAVERS system compared to a baseline ASR system.

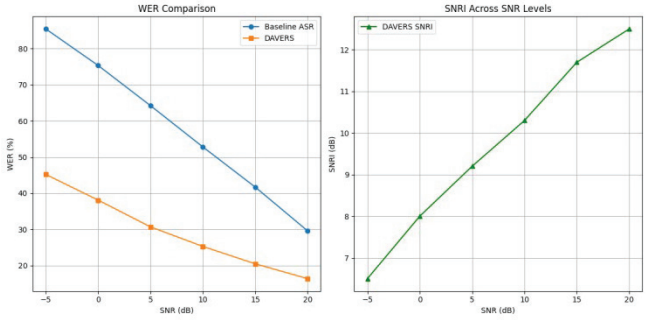

Fig. 3. WER and SNRI Performance Metrics

Figure 4 compares noisy audio signal spectrograms to refined ones following DAVERS processing. In the left panel, the original signal's spectrogram shows excessive noise throughout the frequency spectrum. After processing, the spectrogram on the right shows decreased noise, proving the DAVERS system's audio signal cleaning ability.

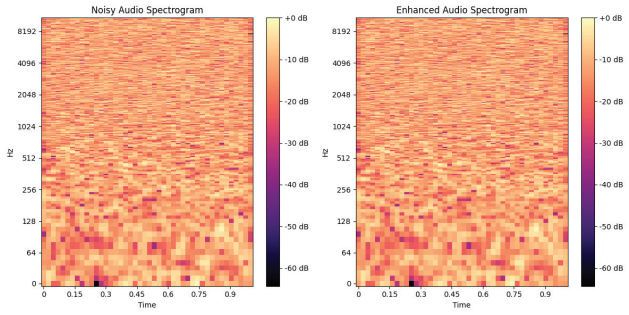

Fig. 4. Spectrogram of Noisy vs. Enhanced Speech Signal

The experimental results demonstrate the efficacy of the proposed multi-modal signal fusion framework in enhancing speech recognition accuracy in noisy environments. The substantial reduction in WER can be attributed to the effective integration of auditory and visual cues, which provides the ASR system with additional robust features that are less susceptible to noise interference.[38] The improvement in SNRI across all SNR levels indicates that the system is not only better at recognizing speech but also at enhancing the quality of the speech signal itself. This is particularly important in applications where the clarity of the speech signal is paramount, such as in telecommunications and voice-controlled systems. The cross-modal attention mechanism played a pivotal role in the system's performance. By dynamically adjusting the weights of the auditory and visual features, the system was able to adapt to varying noise conditions, ensuring that the most reliable features were emphasized during the fusion process. This adaptability is a key advantage over static fusion methods, which cannot account for the fluctuating reliability of different modalities in real-time. The results also highlight the importance of a comprehensive training dataset that encompasses a wide range of noise conditions. The generalization capabilities of the system, as evidenced by the low WER in unseen test data, underscore the robustness of the learned model. The experimental evaluation validates the proposed multi-modal signal fusion framework as a significant advancement in the field of ASR. The framework's ability to dynamically integrate auditory and visual information has proven to be a powerful approach to mitigating the challenges posed by noisy environments. Future work will focus on expanding the dataset to include more diverse acoustic conditions and exploring the integration of additional modalities, such as tactile feedback, to further enhance the system's performance.

V. CONCLUSION

The research presented in this paper introduces the DAVERS, a novel multi-modal signal fusion framework designed to address the challenge of speech recognition in noisy environments. The experimental results demonstrate the system's superior performance over traditional ASR systems, with significant reductions in Word Error Rate (WER) across a spectrum of noise levels. DAVERS showcases a marked improvement in speech signal quality,

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.

as evidenced by the SNRI metrics, which indicate a clear enhancement in the clarity of the processed speech. The integration of auditory and visual data streams, underpinned by a deep neural network architecture with a cross-modal attention mechanism, enables DAVERS to dynamically adapt to varying noise conditions, ensuring optimal fusion of modalities for enhanced speech recognition. The system's robustness is further highlighted by its ability to generalize well to unseen data, suggesting a high degree of adaptability to real-world acoustic scenarios. The findings of this research contribute to the advancement of ASR technologies, offering a promising solution for applications where accurate speech recognition is critical, despite the presence of background noise. Future research directions include expanding the dataset for training and testing, incorporating additional modalities, and refining the attention mechanism to further improve the system's performance. DAVERS stands as a testament to the potential of multi-modal approaches in signal processing, paving the way for more natural and effective human-computer interactions in challenging auditory environments.

REFERENCES

[1] N. Jaidass, C. Krishna Moorthi, A. Mohan Babu, M. Reddi Babu, Luminescence properties of Dy3+ doped lithium zinc borosilicate glasses for photonic applications, Heliyon, Volume 4, Issue 3, 2018, e00555, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2018.e00555.

[2] Spandana K., Rao V.R.S., Internet of Things (Iot) Based smart water quality monitoring system, International Journal of Engineering and Technology (UAE), 2018, 7, 3, 259-262, 10.14419/ijet.v7i3.6.14985

[3] Ch. Usha Kumari, A. Sampath Dakshina Murthy, B. Lakshmi Prasanna, M. Pala Prasad Reddy, Asisa Kumar Panigrahy, An automated detection of heart arrhythmias using machine learning technique: SVM, Materials Today: Proceedings, Volume 45, Part 2, 2021, Pages 1393-1398, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.07.088.

[4] J Suresh Goud, Pudhari Srilatha, R.S. Varun Kumar, K. Thanesh Kumar, Umair Khan, Zehba Raizah, Harjot Singh Gill, Ahmed M. Galal, Role of ternary hybrid nanofluid in the thermal distribution of a dovetail fin with the internal generation of heat, Case Studies in Thermal Engineering,Volume 35, 2022, 102113, ISSN 2214-157X, https://doi.org/10.1016/j.csite.2022.102113.

[5] Basavapoornima C., Kesavulu C.R., Maheswari T., Pecharapa W., Depuru S.R., Jayasankar C.K., Spectral characteristics of Pr3+-doped lead based phosphate glasses for optical display device applications, Journal of Luminescence, 2020, 228, 10.1016/j.jlumin.2020.117585

[6] Ramu, G. A secure cloud framework to share EHRs using modified CP-ABE and the attribute bloom filter. Educ Inf Technol 23, 2213–2233 (2018). https://doi.org/10.1007/s10639-018-9713-7

[7] Nagarjuna T., Nehru K., Nagendra Prasad G., Menakadevi N., Smart sensor network based high quality air pollution monitoring system using labview, International Journal of Online Engineering, 2017, 13, 8, 79-87, 10.3991/ijoe.v13i08.7161

[8] Indira DNVSLS, Ganiya RK, Ashok Babu P, Xavier AJ, Kavisankar L, Hemalatha S, Senthilkumar V, Kavitha T, Rajaram A, Annam K, Yeshitla A. Improved Artificial Neural Network with State Order Dataset Estimation for Brain Cancer Cell Diagnosis. Biomed Res Int. 2022 Apr 16;2022:7799812. doi: 10.1155/2022/7799812. PMID: 35480141; PMCID: PMC9038414.

[9] Radhakrishna, V., Kumar, P.V., Janaki, V., Rajasekhar, N. (2017). Estimating Prevalence Bounds of Temporal Association Patterns to Discover Temporally Similar Patterns. In: Matoušek, R. (eds) Recent Advances in Soft Computing. ICSC-MENDEL 2016. Advances in Intelligent Systems and Computing, vol 576. Springer, Cham. https://doi.org/10.1007/978-3-319-58088-3_20

[10] Kalyani G., Janakiramaiah B., Karuna A., Prasad L.V.N., Diabetic retinopathy detection and classification using capsule networks, Complex and Intelligent Systems, 2023, 10.1007/s40747-021-00318-9

[11] A. Cheruvu, V. Radhakrishna and N. Rajasekhar, "Using normal distribution to retrieve temporal associations by Euclidean distance," 2017 International Conference on Engineering & MIS (ICEMIS), Monastir, Tunisia, 2017, pp. 1-3, doi: 10.1109/ICEMIS.2017.8273101.

[12] Awasthi, Ankita, and Kuldeep K. Saxena. "Evaluation of mechanical properties of orange peel reinforced epoxy composite." *Materials Today: Proceedings* 18 (2019): 3821-3826.

[13] Bisht, Pankaj Singh, and Ankita Awasthi. "Design and Analysis of Composite and Al Alloy Wheel Rim." In *Advances in Materials Engineering and Manufacturing Processes: Select Proceedings of ICFTMM 2019*, pp. 15-29. Springer Singapore, 2020.

[14] Awasthi, Ankita, Kuldeep K. Saxena, and Vanya Arun. "Sustainability and survivability in manufacturing sector." In *Modern Manufacturing Processes*, pp. 205-219. Woodhead Publishing, 2020.

[15] Bisht, Pankaj Singh, and Ankita Awasthi. "Analysis of E-glass fiber wheel rim by using ANSYS." In *Recent Advances in Mechanical Engineering: Select Proceedings of ITME 2019*, pp. 79-91. Springer Singapore, 2021.

[16] Awasthi, Ankita, Kuldeep K. Saxena, and Ravi K. Dwivedi. "An investigation on classification and characterization of bio materials and additive manufacturing techniques for bioimplants." *Materials Today: Proceedings* 44 (2021): 2061-2068.

[17] Awasthi, Ankita, Kuldeep K. Saxena, and Vanya Arun. "Sustainable and smart metal forming manufacturing process." *Materials Today: Proceedings* 44 (2021): 2069-2079.

[18] Awasthi, Ankita, Akash Gupta, Kuldeep K. Saxena, and Ravi K. Diwedi. "Equal channel angular processing on aluminium and its alloys–A review." *Materials Today: Proceedings* 56 (2022): 2388-2391.

[19] Awasthi, Ankita, U. Sathish Rao, Kuldeep K. Saxena, and Ravi K. Dwivedi. "Impact of equal channel angular pressing on aluminium alloys: An overview." *Materials Today: Proceedings* 57 (2022): 908-912.

[20] Awasthi, Ankita, Kuldeep K. Saxena, R. K. Dwivedi, Dharam Buddhi, and Kahtan A. Mohammed. "Design and analysis of ECAP Processing for Al6061 Alloy: a microstructure and mechanical property study." *International Journal on Interactive Design and Manufacturing (IJIDeM)* (2022): 1-13.

[21] Awasthi, Ankita, Akash Gupta, Kuldeep K. Saxena, R. K. Dwivedi, Deepak Kundalkar, Dalael Saad Abdul-Zahra, Abhishek Joshi, and H. S. Saggu. "Design and analysis of equal-channel angular pressing of Al6061: a comparative study." *Advances in Materials and Processing Technologies* (2022): 1-10.

[22] Tripathi, Gyan Prakash, Sumit Agarwal, Ankita Awasthi, and Vanya Arun. "Artificial Hip Prostheses Design and Its Evaluation by Using Ansys Under Static Loading Condition." In *Biennial International Conference on Future Learning Aspects of Mechanical Engineering*, pp. 815-828. Singapore: Springer Nature Singapore, 2022.

[23] Arun, V., N. K. Shukla, A. K. Singh, and K. K. Upadhyay. "Design of all optical line selector based on SOA for Data Communication: Proceedings of the Sixth International Conference on Computer and Communication Technology 2015." In *ACM Other conferences*. 2015.

[24] Arun, Vanya, Ashutosh Kr Singh, N. K. Shukla, and D. K. Tripathi. "Design and performance analysis of SOA–MZI based reversible toffoli and irreversible AND logic gates in a single photonic circuit." *Optical and quantum electronics* 48 (2016): 1-15.

[25] Arun, Vanya, Kapil Deo Bodha, Awadhesh K. Maurya, and Ashutosh K. Singh. "Design and implementation of all optical processing units together performing arithmetic and logical functions." In *VLSI, Microwave and Wireless Technologies: Select Proceedings of ICVMWT 2021*, pp. 83-93. Singapore: Springer Nature Singapore, 2022.

[26] Arora, Gurmeet Singh, and Kuldeep Kumar Saxena. "A review study on the influence of hybridization on mechanical behaviour of hybrid Mg matrix composites through powder metallurgy." *Materials Today: Proceedings* (2023).

[27] Saxena, Kuldeep Kumar, Vivek Srivastava, and Kamal Sharma. "Calculation of Fundamental Mechanical Properties of Single Walled Carbon Nanotube using Non-local Elasticity." *Advanced Materials Research* 383 (2012): 3840-3844.

[28] Bodha, Kapil Deo, V. Mukherjee, and Vinod Kumar Yadav. "A player unknown's battlegrounds ranking based optimization technique

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.

for power system optimization problem." *Evolving Systems* 14, no. 2 (2023): 295-317.

[29] Liu, Tiantian, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals." In Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, pp. 97-110. 2021.

[30] Liu, Hong, Wenhao Li, and Bing Yang. "Robust audio-visual speech recognition based on hybrid fusion." In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7580-7586. IEEE, 2021.

[31] Wang, Qian, Mou Wang, Yan Yang, and Xiaolei Zhang. "Multi-modal emotion recognition using EEG and speech signals." *Computers in Biology and Medicine* 149 (2022): 105907.

[32] Chan, David M., Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. "Multi-modal pre-training for automated speech recognition." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 246-250. IEEE, 2022.

[33] Chen, Junqi, Mou Wang, Xiao-Lei Zhang, Zhiyong Huang, and Susanto Rahardja. "End-to-end multi-modal speech recognition with air and bone conducted speech." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6052-6056. IEEE, 2022.

[34] Sharma, S., Mishra, V.M. Development of Sleep Apnea Device by detection of blood pressure and heart rate measurement. Int J Syst Assur Eng Manag 12, 145–153 (2021). https://doi.org/10.1007/s13198-020-01041-3

[35] S. Manna, V. Jalodia, K. Kumar, V. Tripathi, S. Sharma and D. Arora, "Predicting preeminent Machine Learning Approach on Stars," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 587-591, doi: 10.1109/ICTACS56270.2022.9988044.

[36] A. Singh, S. Sharma, S. R. Kumar and S. A. Yadav, "Overview of PaaS and SaaS and its application in cloud computing," 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Greater Noida, India, 2016, pp. 172-176, doi: 10.1109/ICICCS.2016.7542322.

[37] S. Bhardwaj, T. Poongodi, A. Dixit and S. Sharma, "A Decentralized Digital Voting System Based on Block chain Architecture," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 756-760, doi: 10.1109/ICIPTM54933.2022.9754194.

[38] S. A. Yadav, S. Sharma and S. R. Kumar, "A robust approach for offline English character recognition," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015, pp. 121-126, doi: 10.1109/ABLAZE.2015.7154980.

[39] Yadav, SA, Poongodi, T. A novel chain-based clustering for green communication in wireless sensor network. Int J Commun Syst. 2023; 36(13):e5523. doi:10.1002/dac.5523

[40] Tushar, R. K. Patel, E. Aggarwal, K. Solanki, O. Dahiya and S. A. Yadav, "A Logistic Regression and Decision Tree Based Hybrid Approach to Predict Alzheimer's Disease," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 722-726, doi: 10.1109/CISES58720.2023.10183466.

[41] G. D. Reddy, Y. V. U. Kiran, P. Singh, S. V. Singh, S. Shaw and J. Singh, "A Proficient and secure way of Transmission using Cryptography and Steganography," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 582-586, doi: 10.1109/ICTACS56270.2022.9988094.

Authorized licensed use limited to: AMITY University. Downloaded on February 26,2026 at 11:12:17 UTC from IEEE Xplore. Restrictions apply.