International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Multi-Modal Fusion for Enhanced Image and Speech Recognition in AI Systems

Ankur Tak

Abstract: ***This research investigates the integration of multi-modal information, specifically images and speech, to enhance the recognition capabilities of artificial intelligence (AI) systems. Adopting an interpretive philosophy and employing a deductive approach, the study explores the potential of dynamic attention mechanisms, semi-supervised learning, and cross-domain adaptation techniques. A descriptive research design is employed, utilizing secondary data collection from reputable academic sources. The research critically evaluates the feasibility and applicability of hardware optimization for efficient multi-modal processing, considering factors like specialized processors and parallel computing. The study presents a thorough analysis of dynamic attention mechanisms, emphasizing their role in dynamically allocating attention across different modalities based on contextual relevance. Additionally, it delves into semi-supervised learning techniques, showcasing their ability to leverage both labeled and unlabeled data for improved recognition performance. Cross-domain adaptation techniques are explored to facilitate the seamless deployment of multi-modal fusion models in diverse real-world scenarios.***

Keywords: AI systems, knowledge, connecting, integrating, multi-modal classification, aural, visual information

1. Introduction

1.1 Research background

A key topic for improving machine learning systems is its incorporation of multi-modal input, especially speech and visuals. Image and speech detection have traditionally been viewed as separate jobs [1]. The incorporation of visual and acoustic data, however, has enormous promise for improving the capacities of AI systems in a variety of scenarios. Because it makes use of the complementary nature of both auditory and visual clues, this fusion makes it possible to comprehend complicated real-world surroundings more thoroughly [2]. Artificial intelligence systems can improve their knowledge of context and accuracy in tasks like item identification, scene interpretation, and communication between humans and computers by concurrently integrating these modalities. Despite the potential, there are still difficulties in effectively connecting and integrating these many data sources [3].In order to overcome these obstacles, our research aims to investigate novel approaches and fusion methods ultimately aiding in the creation of more powerful and adaptable AI systems that are capable of continuous multi-modal classification.

1.2 Purpose and goals of the research

Research Goal
The goal of the project is to improve the recognition of speech and images in artificial intelligence systems using efficient multipurpose fusion methods.

Objectives:

- To research and choose the best feature extraction techniques for speech and image data.
- To create and use fusion methods (such as early fusion, late the fusion reaction, or hybrid fusion) for combining voice and visual data.
- Employing comparison datasets to compare the effectiveness of multiple communication integration models to single-modal methods.
- To evaluate the suggested multi-modal fusion approaches' robustness and generalizability in light of diverse practical situations.

1.3 Research Rationale

The potential for improving the abilities of AI systems through the combination of speech and visual data is enormous. Integrating auditory as well as visual knowledge, which are typically handled as independent modalities, offers a more thorough grasp of complicated settings [4]. This fusion takes advantage of the complimentary nature of aural as well as visual information, improving accuracy in tasks like item recognition and scene comprehension. Despite the potential, there are still difficulties in successfully integrating these many data streams [5]. This research fills this gap by looking into fresh approaches and fusion approaches with the goal of advancing the creation of more reliable and adaptable AI systems that are capable of intuitive multi-modal comprehension and have numerous applications in machine learning as well as processing of natural languages, among other areas.

2. Literature Review

2.1 Feature Extraction Methods for Image and Speech Data

A crucial stage in the statistical analysis of language as well as image data is feature extraction, which enables the alteration of unprocessed data into a format that is easier for artificial intelligence (AI) algorithms to understand and analyze [6]. Traditional techniques like Distribution of orientation gradients (HOG) and Scale-Invariant Frequency Transform (SIFT) have been extensively employed in the context of photographs to extract distinguishing features like corners, borders, and textures [7]. Furthermore, state-of-the-art achievement in picture recognition tasks has been achieved using deep learning approaches, particularly neural networks using convolution (CNNs), which have demonstrated extraordinary proficiency in autonomously learning complex characteristics directly from actual pixel

Paper ID: SR231208202748 DOI: https://dx.doi.org/10.21275/SR231208202748

values. Mel-frequency cepstral coefficients (MFCCs) have become a standard in identifying features for speech data [8].

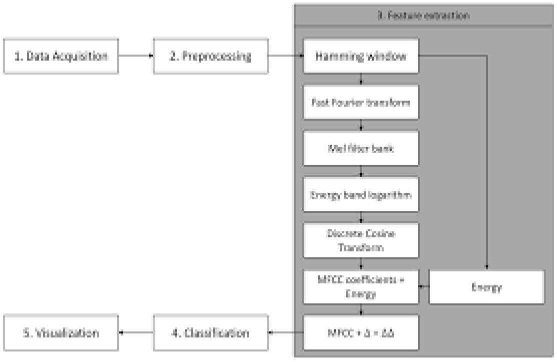

Figure 2.1.1: Feature Extraction Methods Process

MFCCs capture pertinent data for activities like speech by encapsulating the spectral features of signals used for speech [9]. Furthermore, deep learning techniques, such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), have shown significant effectiveness in understanding raw language waveforms without the use of manually created characteristics.

2.2 Multi-modal Fusion Techniques for Image and Speech Recognition

When merging data from several sources, such as audio and images, to improve identification capabilities in artificial intelligence (AI) systems, multi-modal fusion approaches are crucial. In this situation, numerous strategies are used [10].

- Early Fusion: This method combines characteristics at the data entry level before they are fed into the framework. They are extracted from several sources [11]. This method keeps the original data organization intact, enabling cooperative interpretation of both visual and audio input right away.
- Late Fusion: In delayed fusion, the modalities are processed independently using specialized models, and the resultant scores or features are combined at a later time, usually for making decisions [12]. This method enables more in-depth assessment of each modality prior to merging.

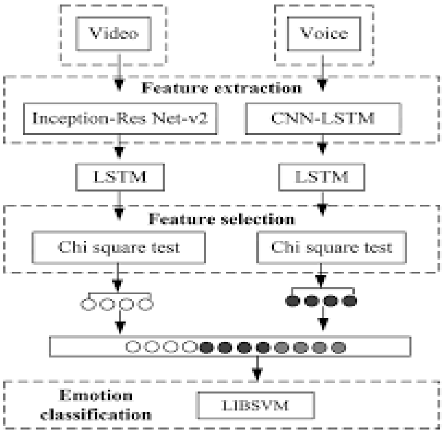

Figure 2.2.1: Multi-modal Fusion Techniques

As the name implies, hybrid fusion incorporates aspects of both late and early fusion. It involves a series of phases that begins with some level of integration, is followed by independent processing, as well as ends with a final fusion step [13].

Attention Mechanisms: Based on their applicability to the job at hand, mechanisms for attention constantly weigh the impact of various modalities [14]. This enhances the model's capacity to adapt to various circumstances by allowing it to concentrate on particular elements of the supplied data.

2.3 Performance Evaluation of Multi-modal Fusion Models

A crucial step in determining the efficiency of multi-modal integration models for improving the recognition of words and images in AI systems is an assessment of performance [15]. To evaluate the effectiveness of each of these models, a variety of measures and approaches are used:

These two essential criteria are used to assess how accurately predictions are made in general. Precision evaluates the reliability of positive predictions, whereas accuracy measures the percentage of correctly categorized instances [16].

F1-Score: The F1-score offers a fair evaluation of a model's recall as well as accuracy. It is very helpful when working with datasets that are unbalanced.

These metrics, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC), are used for the classification of binary data and give information about how well the model recognizes differences between groupings [17].

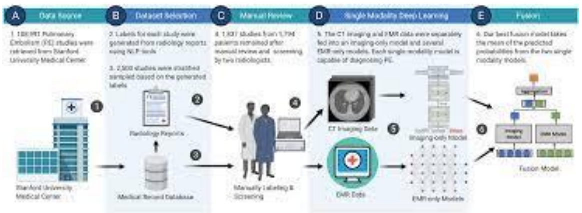

Figure 2.3.1: Multi-modal Fusion Models

- Confusion Matrix: The number of genuine positives, genuine negatives, false positives, and false negatives are displayed throughout this matrix, which offers a thorough evaluation of the model's performance [18].
- Cross-Validation: By educating and evaluating the model on several subsets of the data, approaches like k-fold cross-validation can ensure consistency when performing the assessment [19].
- Comparison with Baseline Models: This illustrates the effectiveness of the fusion approach by comparing the efficiency of multi-modal synthesis models to single-modal systems or other known benchmarks.

2.4 Robustness and Generalization Analysis in Real-world Scenarios

The efficiency of multi-modal fusion models in actual-life situations must be assessed in terms of durability and applicability analysis [20].

- Robustness to Noise and Variability: It's critical to evaluate how effectively the model functions in the absence of noise, changes in lighting, or visual clutter [21]. Even in less-than-ideal situations, solid simulations should maintain outstanding precision.
- Ability to Adjust to Unknown Data: A strong model must be able to apply what it has learned to new types of data. Confusion Matrix: This matrix provides a detailed assessment of the model's accuracy by showing the number of actual positives, true negatives, false positives, and false negatives [22].
- Cross-Validation: Methods like the k-fold cross-validation technique can guarantee consistency while carrying out analyses by training and assessing the model on multiple subsets of the data [23].
- Comparison with Baseline Models: By contrasting the success of multi-modal synthesized models to single-modal methods or other recognized comparisons, this demonstrates the efficacy of the integration approach [24].

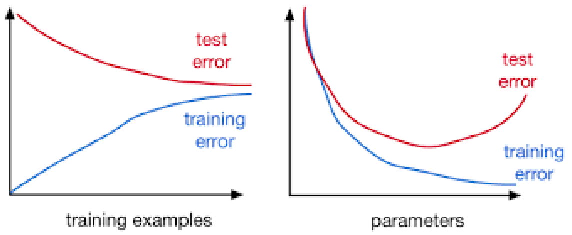

Figure 2.4.1: Regulation in Machine Learning

during a training session. To make sure that the algorithm can handle fresh and varied inputs, this necessitates trying the model with information that was not included in the initial data set [25].

- Transferable Learning Skills: It is vital to look into how much knowledge from one domain or dataset can be used in an alternate but related topic [26]. This evaluates how well the model can use previously learned characteristics to perform better.
- Efficiency and Real-time Processing: Rapid processing is frequently needed in real-world applications. It is crucial to examine the model's computational efficacy and speed, particularly in situations where immediate responses are essential [27].
- Handling of Out-of-Distribution Data: Data that differs considerably from what was used for training may be introduced by real-world circumstances [28]. It is crucial for the algorithm's practical application to assess how well it can identify and manage out-of-distribution inputs.

2.5 Literature Gap

The majority of the currently available material focuses on voice or image-only single-modal identification systems. Effective multipurpose fusion strategies for integrating the two modalities have only received little scientific attention. Additionally, there are very few thorough studies that investigate the reliability and generalization capacities of such models in practical contexts. This study aims to close this gap by offering a comprehensive analysis of multipurpose fusion for improved picture and speech identification.

3. Methodology

The intricate interactions between multi-modal fusion approaches as well as their effects on picture and speech identification in AI systems are explored in this study using an interpretivist perspective [29]. This viewpoint highlights the need of comprehending the meaning that people assign to phenomena while acknowledging the subjective character of human perceptions. Starting with a theoretical framework developed from previous works, the method of deductive reasoning is used [30]. Creating assumptions based on accepted principles and then testing them with direct evidence are involved in this. This study seeks to validate or disprove preexisting notions about multi-modal fusion for improved recognition by using logic of deductive reasoning [31]. To systematically observe, document, and study the traits and behaviours of the multi-modal fusion approaches and their effects, an exploratory approach is adopted. Without interfering with natural environments, this design enables an in-depth knowledge of the present condition and operation of these techniques. The majority of the secondary data included in this study was acquired from credible academic publications, conference proceedings, books, and reliable reports [32]. This involves research on methods for multi-modal fusion, voice and picture recognition, as well as performance assessment measures. Studies and technical articles about artificial intelligence (AI) systems and their potential uses are also consulted. Academic databases including IEEE Xplore, Google Scholar, among others, and PubMed are used in a thorough search. Requests for results are honed using Boolean operators and pertinent keywords (such as "multi-modal fusion," "image recognition," "speech recognition," and "AI systems") [33]. Literature is chosen based on its applicability, reliability, and date of distribution. The list contains only recent, trustworthy reports on technology, peer-reviewed articles, as well as papers presented at conferences. Relevant data is methodically retrieved from chosen sources. This covers information on model designs, feature extraction approaches, fusion strategies, assessment measures, and performance outcomes [34]. To find common patterns, new approaches, and regions of convergence or differentiation within the multi-modal integration field, the data collected is synthesized. To identify the benefits and drawbacks of various fusion procedures, comparative analysis is used. To guarantee both reliability and accuracy, the results are cross-referenced with additional resources [35]. A further assessment of the literature as well as guidance from experts are used to address any anomalies or ambiguities. The project aims to offer useful insights into the effectiveness of multi-modal fusion methods of improving picture and speech detection in AI systems by utilizing this technological technique based on interpretive theory [36]. This method respects the varied perceptions of the phenomena being studied while allowing for a thorough and methodical evaluation of the corpus of available knowledge.

4. Results

4.1 Hardware Optimization for Multi-modal Processing

A customized approach to the development of hardware and usage is required for effective multi-modal processing. Particularly in applications that operate in real time, conventional computing architecture might not fully realize the endless possibilities of multi-modal integration [37]. The strategies for maximizing hardware configurations to satisfy the computing requirements of simultaneous photo and audio processing are covered in this technological

- Specialized Processors and Parallel Computing: Multi-modal integration methods can be executed in parallel by utilizing Tensor Processor Units (TPUs), graphical processing units (GPUs), and other types of specialized computer accelerators. This dramatically increases both the speed and effectiveness of the processes [38]. The project attempts to attain real-time results when performing recognition applications by utilizing the enormous computing capability of these devices.
- Distributed Computing Architectures: Investigating the viability of distributed computation frameworks that can handle massive amounts of multipurpose datasets, including clusters with cloud-based infrastructures, offers scalability with robustness [39]. This strategy makes sure that processing power grows as complicated as the fusion methods used.
- Quantization and Low-Precision Arithmetic: Quantifying and low-precision mathematics are two methods that can be used to minimize memory overall computational needs while improving the use of hardware assets without compromising accuracy [40]. This is especially important in situations involving edge computing or places with limited resources.

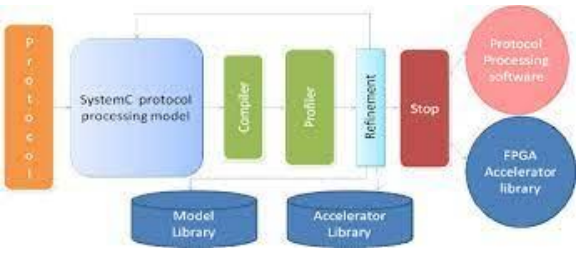

Figure 4.1.1: Optimization for Multi-modal Processing

- Hardware-Software Co-design: Adapting software solutions to hardware architectures enables the smooth integration of efficient algorithms. A symbiotic link between both software and hardware parts is ensured by designing the hardware to work with particular fusion methods and recognition jobs.
- Energy-Efficient Processing: For applications running in locations with limited power sources or battery power, energy-efficient hardware design is essential [41]. To balance both performance and energy consumption, strategies like dynamic electrical frequency and voltage scalability (DVFS) as well as voltage gating are being investigated.

4.2 Dynamic Attention Mechanisms in Fusion Models

Multi-modal merging models can be made more flexible and effective by using dynamic mechanisms for focus. These processes give the model the ability to dynamically assign various levels of attention to distinct modalities by considering the importance of each to the given context [42]. For better picture and understanding of speech in artificial intelligence (AI), this technical area focuses on investigating and developing dynamic attention techniques.

- Contextual Relevance Estimation: Estimating relevance to context Dynamical attention processes determine which modality—speech or image—contributes the greatest amount to the present inference by examining the surroundings of a given activity [43]. This enables the model to allocate resources wisely and concentrate on the inputs that are most instructive.
- Temporal and Spatial Attention: Different jobs could call for various temporal and geographical foci. As an example, in the analysis of video, attention to space focuses on important areas inside a picture, while chronological attention prioritizes pertinent frames.
- Multi-level Attention Hierarchies: The model can fine-tune its focus at various granularities because to the inclusion of numerous levels of concentration hierarchies [44]. This flexibility is especially important for jobs that call for both the highest level and smooth contextual awareness.

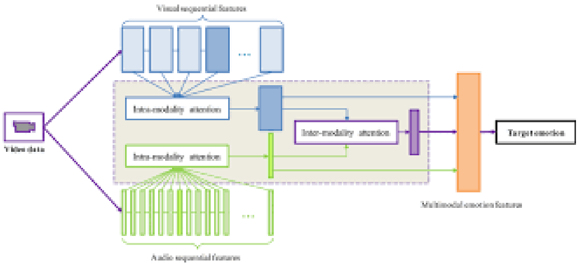

Figure 4 2.1: Attention Mechanisms in Fusion Models

- Attention Fusion Techniques: It is crucial to look at fusion methods that blend ratings of attention from many modalities. Techniques including feature-level combination, in which attention scores are merged before processing continues, may be used in this situation.
- Dynamic Modality Weighting: To avoid overemphasizing on a single source of information in situations when one modality might predominate the others, dynamic mechanisms of focus modify modality weights [45]. This guarantees that both modalities contribute equally.

4.3 Semi-Supervised Learning for Enhanced Recognition

A potent paradigm that uses unlabeled as well as labeled information to enhance recognition ability is partially supervised learning [46]. For improved picture and pronunciation recognition under AI systems, this technical problem investigates the use of semi-supervised learning approaches in a setting of multipurpose fusion.

Leveraging Unlabeled Data: Making Use of Unlabeled Data In practical situations, unlabeled data frequently exceed labeled data by a large margin [47]. This abundance is tapped into by semi-supervised learning, which uses both kinds of data for education. Unlabeled data facilitates the learning of reliable representations for features from a variety of examples.

Self-training and Co-training Techniques: In self-teaching a model is iteratively trained on the data that has been labeled and then used to accurately label further unlabeled samples [48]. Contrarily, co-training entails training several models on various subsets of characteristics or data, then swapping predictions for additional improvement.

Consistency Regularization: This technique requires that forecasts remain accurate even when the input data is slightly distorted [49]. It aids in the model's learning of interpretations that are reliable and transferable to various input changes.

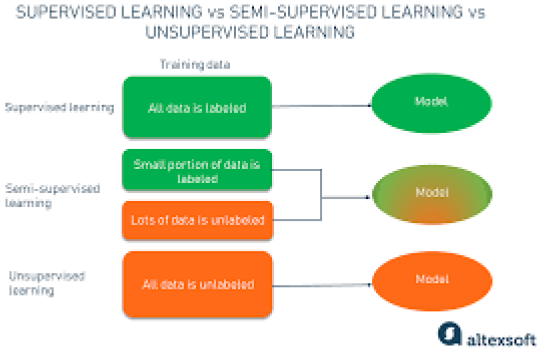

Figure 4.3.1: Semi-Supervised Learning

Pseudo-labeling and Confidence Thresholding: Using the model's assumptions as a guide, pseudo-labeling provides labels to unlabeled information. Confidence thresholding reduces the risk of noisy labeling by ensuring that only those with high-confidence guesses are used for training.

Transfer Learning and Fine-tuning: Large-scale datasets with model training can be effective feature extraction devices [50]. Utilizing prior information is made possible by tweaking the models for the particular recognition challenge at hand.

4.4 Cross-domain Adaptation for Real-world Deployment

Multi-modal fusion models must be used in a variety of real-world contexts that may be very different from the controlled conditions in which they were initially trained. This crucial difficulty is addressed by cross-domain adaptations. This technical subject focuses on methods for modifying models to ensure their reliable operation across many domains and their suitability for usage in real-world situations [51].

- Domain Shift Analysis: It is crucial to comprehend the distinctions between the training environment and the actual use domain. Identifying differences in data distribution, ambient circumstances, and contextual elements that could affect the model's performance is required for this.
- Domain Adaptation Techniques: Models can be more easily adapted to different contexts by using approaches including domain adversarial learning, which reduces the distribution difference between subdomains [52]. Feature space alignment and bias reduction fall under this category.

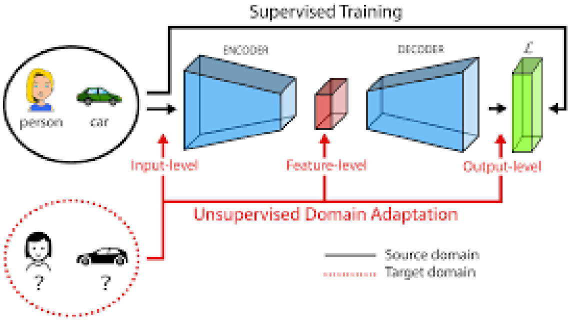

Figure 4.4.1: Domain Adaptation for Real-world Deployment

Data Augmentation and Synthesis: Data enhancement and synthesis methods can be quite useful for creating synthetic data that mimics actual environmental conditions. This might entail methods like style transfer, which modify the style of photographs to fit the intended domain.

Transfer Learning with Fine-tuning: Leveraging previous experience while adjusting for particular deployment conditions is made possible by pre-training systems on data from a source sector and then fine-tuning them based on data from a target domain [53]. This makes it possible for the model to gain from a larger body of knowledge.

Ensemble Methods for Domain Adaptation: The robustness while adaptability about a fusion approach can be improved by combining numerous models that have been trained on various source themes and improving them to fit the domain of interest.

5. Evaluation and Conclusion

5.1 Critical Evaluation

The study "Multi-modal Integration for Enhanced Vision and Voice Recognition with AI Systems" offers a thorough and methodically sound inquiry into a significant field of AI. Deductive reasoning and the adoption by an interpretive approach create a strong theoretical base. The descriptive approach and secondary data collecting are in line with the goals of the study. However, there are several things worth thinking about. While additional sources of information are a treasure of knowledge, the study may be enhanced by using original data collecting since it can yield insights that are unique to the study's setting. The selection from research and its incorporation standards may also generate bias, which must be addressed and minimized. Furthermore, the computer's optimization theme presents a promising approach for improving computing efficiency, but real-world implementation difficulties can appear, particularly in contexts with limited resources. Additionally, the dynamic mechanisms for attention theme needs to be carefully tuned to achieve the ideal balance among adaptability as well as stability, despite its importance.

5.2 Research recommendation

The investigation on "Multi-modal Synthesis for Improved Image as well as Language Recognition across AI Systems" might be improved in a number of ways according to the results and evaluation.

- **Incorporate Primary Data Collection:** It is possible to add context-specific information, real-world viewpoints, and theoretical conclusions to secondary data by supplementing it with original sources.
- **Mitigate Potential Bias in Literature Selection:** Identify any biases in the literature selection process and put procedures in place to ensure an equitable representation of opinions and research methods [54].
- **Conduct Practical Implementations:** To show the viability and efficacy of suggested multi-modal fusion strategies, and validate theories with practical usage and useful experiments.
- **Explore Edge Cases and Resource-Constrained Environments:** Make sure that the research is still usable in a variety of real settings by looking into how well the combination of models can adapt in situations with constrained computing power.

5.3 Future work

Future studies in this discipline should concentrate on a few key areas in order to progress multifaceted fusion for improved picture and language recognition with AI systems.

- **Fine-grained Modality Integration:** Explore sophisticated methods for combining voice and picture data with a finer level of detail, enabling a more nuanced as well as contextually enriched fusion.
- **Continual Learning and Adaptation:** Develop mechanisms for ongoing learning so that models can change flexibly in response to shifting environments and altering recognition problems [55].
- **Cross-modal Transfer Learning:** To increase the generalization of models and effectiveness, look into techniques for applying knowledge learned from one recognized task to yet another regardless of how the modalities are different.
- **Explain ability and Interpretability:** Increase the explain ability as well as interpretability of multipurpose fusion algorithms to make it easier to comprehend how they get to their decisions, especially in important applications.
- **Human-in-the-Loop Fusion:** Integrate user input and subject-matter expertise into the process of fusion to enable a more hands-on and team-based method for tasks related to recognition.

References

[1] P. Bhatt *et al*, "Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions," *Brain Informatics*, vol. 10, *(1)*, pp. 18, 2023. Available: https://www.proquest.com/scholarly-journals/machine-learning-cognitive-behavioral-analysis/docview/2843969816/se-2. DOI: https://doi.org/10.1186/s40708-023-00196-6.

[2] S. Han *et al*, "Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph," *Complex & Intelligent Systems*, vol. 9, *(5)*, pp. 4995-5012, 2023. Available: https://www.proquest.com/scholarly-journals/lightweight-dense-video-captioning-with-cross/docview/2867416769/se-2. DOI: https://doi.org/10.1007/s40747-023-00998-5.

[3] X. Qi *et al*, "Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review," *Frontiers in Aging Neuroscience*, 2023. Available: https://www.proquest.com/scholarly-journals/noninvasive-automatic-detection-alzheimers/docview/2856157707/se-2. DOI: https://doi.org/10.3389/fnagi.2023.1224723.

[4] Z. Ding *et al*, "Adaptive visual–tactile fusion recognition for robotic operation of multi-material system," *Frontiers in Neurorobotics*, 2023. Available: https://www.proquest.com/scholarly-journals/adaptive-visual-tactile-fusion-recognition/docview/2827370627/se-2. DOI: https://doi.org/10.3389/fnbot.2023.1181383.

[5] F. Li *et al*, "GCF2-Net: global-aware cross-modal feature fusion network for speech emotion recognition," *Frontiers in Neuroscience*, 2023. Available: https://www.proquest.com/scholarly-journals/gcf2-net-global-aware-cross-modal-feature-fusion/docview/2808756042/se-2. DOI: https://doi.org/10.3389/fnins.2023.1183132.

[6] Z. Liu *et al*, "Deep learning based brain tumor segmentation: a survey," *Complex & Intelligent Systems*, vol. 9, *(1)*, pp. 1001-1026, 2023. Available: https://www.proquest.com/scholarly-journals/deep-learning-based-brain-tumor-segmentation/docview/2778776786/se-2. DOI: https://doi.org/10.1007/s40747-022-00815-5.

[7] C. Kaur *et al*, "Integrating Transfer Learning and Deep Neural Networks for Accurate Medical Disease Diagnosis from Multi-Modal Data," *International Journal of Advanced Computer Science and Applications*, vol. 14, *(8)*, 2023. Available: https://www.proquest.com/scholarly-journals/integrating-transfer-learning-deep-neural/docview/2869803778/se-2. DOI: https://doi.org/10.14569/IJACSA.2023.0140857.

[8] I. Nazir *et al*, "Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion," *Journal of Sensors*, vol. 2023, 2023. Available: https://www.proquest.com/scholarly-journals/machine-learning-based-lung-cancer-detection/docview/2857678396/se-2. DOI: https://doi.org/10.1155/2023/6683438.

[9] Andrada-Livia Cîrneanu, D. Popescu and D. Iordache, "New Trends in Emotion Recognition Using Image Analysis by Neural Networks, A Systematic Review," *Sensors*, vol. 23, *(16)*, pp. 7092, 2023. Available: https://www.proquest.com/scholarly-journals/new-trends-emotion-recognition-using-image/docview/2857446834/se-2. DOI: https://doi.org/10.3390/s23167092.

[10] L. De Rosa *et al*, "Applications of Deep Learning Algorithms to Ultrasound Imaging Analysis in Preclinical Studies on In Vivo Animals," *Life*, vol. 13,

(8), pp. 1759, 2023. Available: https://www.proquest.com/scholarly-journals/applications-deep-learning-algorithms-ultrasound/docview/2857109676/se-2. DOI: https://doi.org/10.3390/life13081759.

[11] H. Quan *et al*, "Big Data and AI-Driven Product Design: A Survey," *Applied Sciences*, vol. 13, *(16)*, pp. 9433, 2023. Available: https://www.proquest.com/scholarly-journals/big-data-ai-driven-product-design-survey/docview/2856797620/se-2. DOI: https://doi.org/10.3390/app13169433.

[12] O. E. Gannour *et al*, "Enhancing Skin Diseases Classification Through Dual Ensemble Learning and Pre-trained CNNs," *International Journal of Advanced Computer Science and Applications*, vol. 14, *(6)*, 2023. Available: https://www.proquest.com/scholarly-journals/enhancing-skin-diseases-classification-through/docview/2843254896/se-2. DOI: https://doi.org/10.14569/IJACSA.2023.0140647.

[13] I. Pulatov *et al*, "Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders," *Sensors*, vol. 23, *(14)*, pp. 6640, 2023. Available: https://www.proquest.com/scholarly-journals/enhancing-speech-emotion-recognition-using-dual/docview/2843126043/se-2. DOI: https://doi.org/10.3390/s23146640.

[14] C. Jiao *et al*, "Contrast-Enhanced Liver Magnetic Resonance Image Synthesis Using Gradient Regularized Multi-Modal Multi-Discrimination Sparse Attention Fusion GAN," *Cancers*, vol. 15, *(14)*, pp. 3544, 2023. Available: https://www.proquest.com/scholarly-journals/contrast-enhanced-liver-magnetic-resonance-image/docview/2843043089/se-2. DOI: https://doi.org/10.3390/cancers15143544.

[15] X. Jiang *et al*, "Deep Learning for Medical Image-Based Cancer Diagnosis," *Cancers*, vol. 15, *(14)*, pp. 3608, 2023. Available: https://www.proquest.com/scholarly-journals/deep-learning-medical-image-based-cancer/docview/2843041919/se-2. DOI: https://doi.org/10.3390/cancers15143608.

[16] R. Ullah *et al*, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, *(13)*, pp. 6212, 2023. Available: https://www.proquest.com/scholarly-journals/speech-emotion-recognition-using-convolution/docview/2836492932/se-2. DOI: https://doi.org/10.3390/s23136212.

[17] S. K. Paul, M. Nicolescu and M. Nicolescu, "Enhancing Human–Robot Collaboration through a Multi-Module Interaction Framework with Sensor Fusion: Object Recognition, Verbal Communication, User of Interest Detection, Gesture and Gaze Recognition," *Sensors*, vol. 23, *(13)*, pp. 5798, 2023. Available: https://www.proquest.com/scholarly-journals/enhancing-human-robot-collaboration-through-multi/docview/2836484635/se-2. DOI: https://doi.org/10.3390/s23135798.

[18] P. Kumar, S. Khalid and H. S. Kim, "Prognostics and Health Management of Rotating Machinery of Industrial Robot with Deep Learning Applications—A Review," *Mathematics*, vol. 11, *(13)*, pp. 3008, 2023. Available: https://www.proquest.com/scholarly-journals/prognostics-health-management-rotating-machinery/docview/2836420369/se-2. DOI: https://doi.org/10.3390/math11133008.

[19] Z. Zhao *et al*, "Lightweight Infrared and Visible Image Fusion via Adaptive DenseNet with Knowledge Distillation," *Electronics*, vol. 12, *(13)*, pp. 2773, 2023. Available: https://www.proquest.com/scholarly-journals/lightweight-infrared-visible-image-fusion-via/docview/2836309198/se-2. DOI: https://doi.org/10.3390/electronics12132773.

[20] D. Mamieva *et al*, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, *(12)*, pp. 5475, 2023. Available: https://www.proquest.com/scholarly-journals/multimodal-emotion-detection-via-attention-based/docview/2829878107/se-2. DOI: https://doi.org/10.3390/s23125475.

[21] J. Zong *et al*, "FCAN–XGBoost: A Novel Hybrid Model for EEG Emotion Recognition," *Sensors*, vol. 23, *(12)*, pp. 5680, 2023. Available: https://www.proquest.com/scholarly-journals/fcan-xgboost-novel-hybrid-model-eeg-emotion/docview/2829876442/se-2. DOI: https://doi.org/10.3390/s23125680.

[22] A. M. Abubakar *et al*, "Deep Neural Networks for Spatial-Temporal Cyber-Physical Systems: A Survey," *Future Internet*, vol. 15, *(6)*, pp. 199, 2023. Available: https://www.proquest.com/scholarly-journals/deep-neural-networks-spatial-temporal-cyber/docview/2829802943/se-2. DOI: https://doi.org/10.3390/fi15060199.

[23] D. Xu, X. Fan and W. Gao, "Multiscale Attention Fusion for Depth Map Super-Resolution Generative Adversarial Networks," *Entropy*, vol. 25, *(6)*, pp. 836, 2023. Available: https://www.proquest.com/scholarly-journals/multiscale-attention-fusion-depth-map-super/docview/2829796325/se-2. DOI: https://doi.org/10.3390/e25060836.

[24] A. Alam, S. Urooj and A. Q. Ansari, "Design and Development of a Non-Contact ECG-Based Human Emotion Recognition System Using SVM and RF Classifiers," *Diagnostics*, vol. 13, *(12)*, pp. 2097, 2023. Available: https://www.proquest.com/scholarly-journals/design-development-non-contact-ecg-based-human/docview/2829793996/se-2. DOI: https://doi.org/10.3390/diagnostics13122097.

[25] Y. Jin, "Facial Gesture Detection using Ubiquitous Acoustic Sensing." Order No. 30493160, State University of New York at Buffalo, United States -- New York, 2023.

[26] J. S. Joseph and C. Lakshmi, "A Novel Framework for Semi-supervised Multiple-label Image Classification using Multi-stage CNN and Visual Attention Mechanism," *International Journal of Advanced Computer Science and Applications*, vol. 14, *(4)*, 2023. Available: https://www.proquest.com/scholarly-journals/novel-framework-semi-supervised-multiple-label/docview/2819915891/se-2. DOI: https://doi.org/10.14569/IJACSA.2023.0140454.

[27] H. A. Younis *et al*, "Multimodal Age and Gender Estimation for Adaptive Human-Robot Interaction: A Systematic Literature Review," *Processes*, vol. 11, *(5)*, pp. 1488, 2023. Available: https://www.proquest.com/scholarly-journals/multimodal-age-gender-estimation-adaptive-human/docview/2819482255/se-2. DOI: https://doi.org/10.3390/pr11051488.

[28] M. Lukac *et al*, "Study on emotion recognition bias in different regional groups," *Scientific Reports (Nature Publisher Group)*, vol. 13, *(1)*, pp. 8414, 2023. Available: https://www.proquest.com/scholarly-journals/study-on-emotion-recognition-bias-different/docview/2818596652/se-2. DOI: https://doi.org/10.1038/s41598-023-34932-z.

[29] M. A. Razzaq *et al*, "A Hybrid Multimodal Emotion Recognition Framework for UX Evaluation Using Generalized Mixture Functions," *Sensors*, vol. 23, *(9)*, pp. 4373, 2023. Available: https://www.proquest.com/scholarly-journals/hybrid-multimodal-emotion-recognition-framework/docview/2812735709/se-2. DOI: https://doi.org/10.3390/s23094373.

[30] A. I. Henry *et al*, "Analyzing Factors Influencing Situation Awareness in Autonomous Vehicles—A Survey," *Sensors*, vol. 23, *(8)*, pp. 4075, 2023. Available: https://www.proquest.com/scholarly-journals/analyzing-factors-influencing-situation-awareness/docview/2806590877/se-2. DOI: https://doi.org/10.3390/s23084075.

[31] S. V. Jitender *et al*, "A Multimodal Feature Fusion Framework for Sleep-Deprived Fatigue Detection to Prevent Accidents," *Sensors*, vol. 23, *(8)*, pp. 4129, 2023. Available: https://www.proquest.com/scholarly-journals/multimodal-feature-fusion-framework-sleep/docview/2806587299/se-2. DOI: https://doi.org/10.3390/s23084129.

[32] S. Borna *et al*, "A Review of Voice-Based Pain Detection in Adults Using Artificial Intelligence," *Bioengineering*, vol. 10, *(4)*, pp. 500, 2023. Available: https://www.proquest.com/scholarly-journals/review-voice-based-pain-detection-adults-using/docview/2806468300/se-2. DOI: https://doi.org/10.3390/bioengineering10040500.

[33] P. Priyadarshinee *et al*, "Alzheimer’s Dementia Speech (Audio vs. Text): Multi-Modal Machine Learning at High vs. Low Resolution," *Applied Sciences*, vol. 13, *(7)*, pp. 4244, 2023. Available: https://www.proquest.com/scholarly-journals/alzheimer-s-dementia-speech-audio-vs-text-multi/docview/2799597721/se-2. DOI: https://doi.org/10.3390/app13074244.

[34] C. Surianarayanan *et al*, "Convergence of Artificial Intelligence and Neuroscience towards the Diagnosis of Neurological Disorders—A Scoping Review," *Sensors*, vol. 23, *(6)*, pp. 3062, 2023. Available: https://www.proquest.com/scholarly-journals/convergence-artificial-intelligence-neuroscience/docview/2791739513/se-2. DOI: https://doi.org/10.3390/s23063062.

[35] X. Li *et al*, "Polarimetric Imaging via Deep Learning: A Review," *Remote Sensing*, vol. 15, *(6)*, pp. 1540, 2023. Available: https://www.proquest.com/scholarly-journals/polarimetric-imaging-via-deep-learning-review/docview/2791713834/se-2. DOI: https://doi.org/10.3390/rs15061540.

[36] J. Singh, B. S. Lakshmi and O. Faust, "Speech Emotion Recognition Using Attention Model," *International Journal of Environmental Research and Public Health*, vol. 20, *(6)*, pp. 5140, 2023. Available: https://www.proquest.com/scholarly-journals/speech-emotion-recognition-using-attention-model/docview/2791654654/se-2. DOI: https://doi.org/10.3390/ijerph20065140.

[37] S. Khairnar *et al*, "Face Liveness Detection Using Artificial Intelligence Techniques: A Systematic Literature Review and Future Directions," *Big Data and Cognitive Computing*, vol. 7, *(1)*, pp. 37, 2023. Available: https://www.proquest.com/scholarly-journals/face-liveness-detection-using-artificial/docview/2791570909/se-2. DOI: https://doi.org/10.3390/bdcc7010037.

[38] Y. Cai, X. Li and J. Li, "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review," *Sensors*, vol. 23, *(5)*, pp. 2455, 2023. Available: https://www.proquest.com/scholarly-journals/emotion-recognition-using-different-sensors/docview/2785240514/se-2. DOI: https://doi.org/10.3390/s23052455.

[39] A. M. Ali *et al*, "Vision Transformers in Image Restoration: A Survey," *Sensors*, vol. 23, *(5)*, pp. 2385, 2023. Available: https://www.proquest.com/scholarly-journals/vision-transformers-image-restoration-survey/docview/2785236762/se-2. DOI: https://doi.org/10.3390/s23052385.

[40] V. Guimarães *et al*, "A Review of Recent Advances and Challenges in Grocery Label Detection and Recognition," *Applied Sciences*, vol. 13, *(5)*, pp. 2871, 2023. Available: https://www.proquest.com/scholarly-journals/review-recent-advances-challenges-grocery-label/docview/2785180584/se-2. DOI: https://doi.org/10.3390/app13052871.

[41] M. Farah, M. Hussain and H. Aboalsamh, "A Bimodal Emotion Recognition Approach through the Fusion of Electroencephalography and Facial Sequences," *Diagnostics*, vol. 13, *(5)*, pp. 977, 2023. Available: https://www.proquest.com/scholarly-journals/bimodal-emotion-recognition-approach-through/docview/2785178801/se-2. DOI: https://doi.org/10.3390/diagnostics13050977.

[42] M. J. Er *et al*, "Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review," *Sensors*, vol. 23, *(4)*, pp. 1990, 2023. Available: https://www.proquest.com/scholarly-journals/research-challenges-recent-advances-popular/docview/2779682259/se-2. DOI: https://doi.org/10.3390/s23041990.

[43] C. Yu *et al*, "Improvement of Acoustic Models Fused with Lip Visual Information for Low-Resource Speech," *Sensors*, vol. 23, *(4)*, pp. 2071, 2023. Available: https://www.proquest.com/scholarly-journals/improvement-acoustic-models-fused-with-lip-visual/docview/2779654488/se-2. DOI: https://doi.org/10.3390/s23042071.

[44] D. Ryumin, D. Ivanko and E. Ryumina, "Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices," *Sensors*, vol. 23, *(4)*, pp. 2284, 2023. Available: https://www.proquest.com/scholarly-journals/audio-visual-speech-gesture-recognition-sensors/docview/2779550153/se-2. DOI: https://doi.org/10.3390/s23042284.

[45] F. Liu and J. Fang, "Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition," *Future Internet*, vol. 15, *(2)*, pp. 65, 2023. Available: https://www.proquest.com/scholarly-journals/multi-scale-audio-spectrogram-transformer/docview/2779526167/se-2. DOI: https://doi.org/10.3390/fi15020065.

[46] M. Mukhiddinov *et al*, "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People," *Sensors*, vol. 23, *(3)*, pp. 1080, 2023. Available: https://www.proquest.com/scholarly-journals/masked-face-emotion-recognition-based-on-facial/docview/2774977976/se-2. DOI: https://doi.org/10.3390/s23031080.

[47] Vasile-Daniel Păvăloaia and G. Husac, "Tracking Unauthorized Access Using Machine Learning and PCA for Face Recognition Developments," *Information*, vol. 14, *(1)*, pp. 25, 2023. Available: https://www.proquest.com/scholarly-journals/tracking-unauthorized-access-using-machine/docview/2767221061/se-2. DOI: https://doi.org/10.3390/info14010025.

[48] M. H. Md *et al*, "Review on the Evaluation and Development of Artificial Intelligence for COVID-19 Containment," *Sensors*, vol. 23, *(1)*, pp. 527, 2023. Available: https://www.proquest.com/scholarly-journals/review-on-evaluation-development-artificial/docview/2761206646/se-2. DOI: https://doi.org/10.3390/s23010527.

[49] A. Kline *et al*, "Multimodal machine learning in precision health: A scoping review," *NPJ Digital Medicine*, vol. 5, *(1)*, 2022. Available: https://www.proquest.com/scholarly-journals/multimodal-machine-learning-precision-health/docview/2732927616/se-2. DOI: https://doi.org/10.1038/s41746-022-00712-8.

[50] B. Fu and J. Sui, "Multi-modal affine fusion network for social media rumor detection," *PeerJ Computer Science*, 2022. Available: https://www.proquest.com/scholarly-journals/multi-modal-affine-fusion-network-social-media/docview/2658983269/se-2. DOI: https://doi.org/10.7717/peerj-cs.928.

[51] Y. Yu, Q. Dong and Y. Ruiteng, "A multi-modal and multi-scale emotion-enhanced inference model based on fuzzy recognition," *Complex & Intelligent Systems*, vol. 8, *(2)*, pp. 1071-1084, 2022. Available: https://www.proquest.com/scholarly-journals/multi-modal-scale-emotion-enhanced-inference/docview/2656974947/se-2. DOI: https://doi.org/10.1007/s40747-021-00579-4.

[52] S. R. Kshirsagar, "Affective Human-Machine Interfaces: Towards Multi-Lingual, Environment-Robust Emotion Detection from Speech." Order No. 30468582, Institut National de la Recherche Scientifique (Canada), Canada -- Quebec, CA, 2022.

[53] G. M. Dimitri, "A Short Survey on Deep Learning for Multimodal Integration: Applications, Future Perspectives and Challenges," *Computers*, vol. 11, *(11)*, pp. 163, 2022. Available: https://www.proquest.com/scholarly-journals/short-survey-on-deep-learning-multimodal/docview/2748270866/se-2. DOI: https://doi.org/10.3390/computers11110163.

[54] Z. Lv *et al*, "Deep Learning for Intelligent Human–Computer Interaction," *Applied Sciences*, vol. 12, *(22)*, pp. 11457, 2022. Available: https://www.proquest.com/scholarly-journals/deep-learning-intelligent-human-computer/docview/2739421901/se-2. DOI: https://doi.org/10.3390/app122211457.

[55] Y. Cui *et al*, "Artificial Intelligence in Spinal Imaging: Current Status and Future Directions," *International Journal of Environmental Research and Public Health*, vol. 19, *(18)*, pp. 11708, 2022. Available: https://www.proquest.com/scholarly-journals/artificial-intelligence-spinal-imaging-current/docview/2716550818/se-2. DOI: https://doi.org/10.3390/ijerph191811708.