Online ISSN 3007-3197 Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 6 (2025)

Advancing Multi-Modal Machine Learning in Smart Environments: Integrating Visual, Auditory, and Sensorial Data for Context-Aware Human–AI Interaction

[1]Engr.Dr. Shamim Akhtar, [2]Ijaz Khan, [3]Amjad Jumani, [4]Mir Rahib Hussain Talpur, [5]Arshad Iqbal, [6]Adnan Ahmed Rafique

Article Details

Keywords: Context awareness, human–AI interaction, multi-modal machine learning,sensor data fusion, smart environments, transformer models

Engr.Dr. Shamim Akhtar
Assistant Professor, Faculty of Engineering Science and Technology, IQRA University, Karachi. shamim.akhtar@iqra.edu.pk
Ijaz Khan
Department of Avionics Engineering, College of Aeronautical Engineering (CAE), National University of sciences and Technology (NUST). ikhan@cae.nust.edu.pk
Amjad Jumani
Lecture at Computer Science Department Bahria university Karachi Campus. amjadjumani1991@gmail.com
Mir Rahib Hussain Talpur
Dept: Information Technology Centre, Uni: Sindh Agriculture University Tandojam
rahibtalpur@gmail.com
Arshad Iqbal
Department of Computer Science, Khushal Khan Khattak Univversity Karak.
arshadktk.uop@gmail.com
Adnan Ahmed Rafique
Assistant Professor, Department of CS and IT, University of Poonch Rawalakot,
adnanrafique@upr.edu.pk

ABSTRACT

As smart environments spread fast, it is necessary to create intelligent systems that can analyze people's actions through many types of data. The study looks into the progress of multi-modal machine learning (ML) by letting machines use visual, auditory and sensory inputs to work with humans in various contexts. The purpose of this study is to contrast the performance of unimodal, early fusion and transformer-based architectures in a smart environment context. Mixed-methods were used to design the study. To train the models, both audio-visual event data and sensor-based activity data were used quantitatively. In the lab, researcher watched users as they interacted with devices in the smart home environment. A crossmodal attention mechanism was used in the transformer architecture to ensure the semantic and temporal alignment of the different inputs. For every model, researcher checked how accurate, precise, recall was, along with the F1-score, latency and user experience. It was found that the transformer-based model did better than the others on all metrics, scoring 89.5% on F1-score and having the lowest latency at 95 milliseconds. The differences were found to be statistically significant by both ANOVA and Tukey's HSD tests ($p < .001$). Most users agreed that the transformer-based system built more trust and satisfaction in them. Therefore, adding multi-modal data to transformer models greatly enhances the speed and intelligence of smart environments. It is necessary for future studies to develop models that perform well in open environments, are efficient for edge devices and follow ethical guidelines.

INTRODUCTION

With AI moving ahead rapidly, machines have the ability to see and understand things in ways that are very like human thinking. At its core, today's advancement relies on multi-modal machine learning (ML) for systems to handle and merge different types of data together to help them respond more appropriately and sensitively. In smart environments, linking varied technologies together for better user experiences, it is important to use multimodal ML. When combining data from many kinds of sensors, these environments gain a complete picture of people's actions and likes which makes AI respond and interact more naturally.

Because it can be used in healthcare,education, transportation and home automation, multi-modal machine learning is strongly connected to smart environments. If researcher bring together images from tests, recordings of doctor-patient conversations and data from wearable devices in healthcare, it can help doctors make better diagnoses and suggest the best treatment options (Wright et al., 2018). Analyzing these observations in educational contexts can let teachers give each student the right kind of experience (Wu, 2024).

While there are many promising uses for multi-modal ML, several obstacles make it hard to use in smart environments smoothly. Among these challenges are the challenge of integrating different kinds of data, the requirement for powerful computing, worries about data privacy and a lack of common methods for data gathering and use (Kiros et al., 2015). Dealing with these difficulties is necessary to support the improvement of context-aware human–AI interactions.In this study, researcher want to look at the current use of multiple ML technologies in smart environments, find the existing challenges and suggest approaches to improve how people interact with AI devices through the use of visual, auditory and sensory data.

RESEARCH BACKGROUND

This means AI systems use various forms of sensory data such as text, pictures, sound and sensor information, to better understand their environment (Baltrušaitis et al., 2019). It is different from classic ML models that process only one kind of data and are therefore less flexible and have fewer ways to interpret context. Due to multi-modal ML, smart environments can develop systems that can notice and react to a range of human actions and changes around them. Smart homes, in particular, make adjustments to lights, temperature and security when data from cameras, microphones and sensors is integrated and this is based on what the occupants want (Polo-Rodríguez et al., 2025). Autonomous vehicles in transportation rely on

multi-modal ML to read traffic situations, notice safety risks and respond right away to passengers' needs (Hori & Hori, 2020).

AI systems can now work with different types of data more effectively due to recent progress in many language models. Using such models helps AIs produce contextual responses that enhance the standard of human–AI exchanges (Liu et al., 2025).On the other hand, using different ML technologies in smart environments is not free from difficulties. It is necessary to use advanced merging methods that can bring together and explain the different types of data found in heterogeneous data sources. Further, using multi-modal data requires high-power computers, so finding good algorithms and improving hardware is necessary. There are worries about user privacy and data safety whenever personal information is collected and worked on. Additionally, because there are no uniform rules for data collection and processing, devices and systems find it difficult to connect with one another (Binns, 2018).Handling these difficulties is important for making multi-modal ML successful in smart environments and letting human–AI interactions respond to the environment.

RESEARCH PROBLEM

Bringing multi-modal machine learning (ML) into smart environments continues to be a challenging task, even with improvements in artificial intelligence. Even though using visual, audio and sensory data can make systems very responsive, most existing models have difficulty bringing together different types of input. Using different kinds of imaging data together is hard which is why it usually takes powerful, advanced algorithms and huge computing resources, even in applications that work in real time (Baltrušaitis et al., 2019; Liu et al., 2025). In addition, issues about protecting privacy and security make it difficult to use smart home and similar systems, mainly when they depend on always watching and sensing people's homes, schools and hospitals (Binns, 2018). Because standardized frameworks and integration protocols are lacking, it becomes much more difficult to scale and link platforms together (Polo-Rodríguez et al., 2025). To overcome these gaps, special attention is needed to close these technical, ethical and infrastructural challenges for intelligent and human-centric interaction in smart environments. While these problems are unresolved, the goal of having AI that really makes life better every day stays elusive.

RESEARCH OBJECTIVES

1. To study the present status of using multi-modal learning in smart environments, examining the approaches and technologies used for collecting and mixing visual, audio

and sensory data.

2. To figure out what problems and obstacles are present when use multi-modal ML in smart environments concerning technology, ethics and standardization.
3. To come up with plans and structures to overcome these issues, with the goal to make multi-modal data combination more effective in human-AI scenarios.

RESEARCH QUESTIONS

Q1. How are researchers currently using technologies and processes to assemble multi-modal ML in modern surroundings?

Q2. What are the leading obstacles and limits that stop multi-modal ML from being fully integrated?

Q3. How can organizations meet these obstacles by coming up with innovative strategies and frameworks?

Q4. Why effective integrations of multi-modal ML techniques affect both the quality and the efficiency of human–AI communication?

SIGNIFICANCE OF THE STUDY

The study is important because it targets the gap that happens between what technology can do today and what people actually need. The study looks at mixing visual, auditory and sensory data to help craft systems that use technology efficiently and respond to users' mental and emotional needs (Arjunan, 2024; Liu et al., 2025). Furthermore, by putting forward answers to important problems like privacy, limited computation and a lack of standards, the article backs the creation of ethical, expandable and compatible AI infrastructures. Studies in those areas can shape new solutions for healthcare (in patient care), education (in learning) and ambient assisted living (for elders), with adaptive AI increasing safety, ease of use and well-being.

LITERATURE REVIEW

THE EMERGENCE OF MULTI-MODAL MACHINE LEARNING

Because of multi-modal ML, AI systems can analyze different kinds of information such as text, audio, video and sensor inputs. While one-sensory models only deal with one kind of input, multi-modal models combine various types of inputs in the same way that researcher humans process them (Baltrušaitis et al., 2019). The usefulness of this method is especially clear in automated homes, healthcare settings and self-driving vehicles, as AI needs to handle challenging data to work well with users (Liu et al., 2025). Moving from unimodal to multi-modal systems represents a major evolution in machine perception and response to their

environment.

CORE TECHNIQUES IN MULTI-MODAL INTEGRATION

The joining of various input streams is mostly possible because of cutting-edge computational methods. Fusion strategies are separated into early fusion, late fusion and both at the same time (hybrid approaches). In early fusion, raw information from various types of sensors is processed together before extracting labels, yet for late fusion, individual models generate labels and these are then combined (Baltrušaitis et al., 2019). In the past few years, transformer architectures have been added to better unite different sources of data. According to Nguyen et al. (2025), MultiTSF is a transformer framework that successfully performs action recognition for humans by using several sensor views at the same time. RAG methods are also succeeding more and more in smart dialogue systems by mixing in instant contextual data (Agrawal, 2025). As a result, researchers are paying more attention to models that are both reliable and can use lesser memory.

APPLICATIONS IN SMART ENVIRONMENTS

Multi-modal ML has played a major role in transforming systems used for diagnostics and monitoring. Clinicians can now make informed decisions by looking at speech, visual equipment reports and data from sensors together (Wright et al., 2018). Statistical experts at the NYU Astrophysics Lab have designed DeepHeart to spot future cardiovascular problems using combined information from your devices, speech and actions. They improve the accuracy of care, allow constant supervision and tailor treatments, mainly for elderly and remote patients. ML is used in smart classrooms by recognizing speech, analyzing facial expressions and observing gestures to see how well students are participating and understanding what is being taught. These systems enable instructional resources to be adapted according to the requirements of the students. Polo-Rodríguez et al. (2025) explained that chatbots that gather context and use multi-modal data together with LLMs improved student engagement and helped them perform better. When sensory data is united, people learn about emotions and learn more actively.

Self-driving cars and systems that manage traffic all use LiDAR, camera footage, voice alerts and GPS data together. In 2020, Hori and Hori introduced a model that joins visual and audio features to find road hazards. This kind of system is needed for safety, particularly in large cities where it's easy for regular systems to fail.Using visual recognition, voice phrases and sensors lets smart homes give a customized experience to each user. New features in

Google Nest and Amazon Alexa allow them to adapt environmental and safety settings by analyzing our talks and movements, in addition to our voice commands. Because of these developments, smart homes are getting ahead of problems instead of reacting to them.

CHALLENGES IN MULTI-MODAL MACHINE LEARNING

Multi-modal ML continues to encounter a range of persistent difficulties. It is still very difficult to bring together and make compatible data that comes from different sources. When data has sensor drift, different frequencies or misaligned timestamps, the results can be of lower quality (Ektefaie et al., 2023). In addition, handling these large amounts of data takes a lot of computing power which most edge devices can't provide. Even though cloud computing is a possible answer, it still causes delays and is unstable if the internet connection isn't solid (Dosovitskiy & Brox, 2016).

Privacy of data continues to be a top priority. Using multi-modal data in places like healthcare and our homes raises a number of ethical questions. According to Binns (2018), AI systems using personal data need to prioritize being fair and clear which they should always strive to be. Making products uniform across all countries is still difficult. If protocols are not used as guidelines by all vendors, increasing the number of users and platforms becomes difficult which limits scalability (Baltrušaitis et al., 2019).

THE FUTURE OF COGNITIVE ROBOTICS

Explainable multi-modal models have gained importance in recent research to improve transparency and trust. The article explains explainability approaches by Liu et al. (2024), grouping them into steps before the model, in the model and after the model which can all support more interpretable outputs in multi-modal ML. Several multi-modal systems now use facial expressions, how the voice sounds and biometrics to figure out users' feelings (Krzemińska, 2025). As a result, users are likely to be more satisfied and involved, mainly in applications for education and mental health.

Context-aware systems are being used more and more. The framework suggested by Zhang et al. (2024) allows multi-agent systems to adjust quickly to shifts in both user actions and the surroundings. Researcher surpass simple rules, instead providing individual services on the go. Furthermore, scientists are developing light versions of multi-modal models that run without difficulty using mobile and IoT gadgets, allowing AI to help in more places than data centers (Agrawal, 2025).

RESEARCH METHODOLOGY

RESEARCH DESIGN

The research approach used here was mixed-methods exploratory, so both experimentation and user feedback were used to study how multi-modal ML models can be used in smart environments. The numerical part deals with training ML models using sight, sound and environmental information and the human part studies user feedback with the prototype smart setting. With this arrangement, the research covers both how the system runs and how it interacts with people. The mixed method is very effective for evaluating both the technical aspects of systems and the way users see them, providing a complete explanation of systems in action (Creswell & Plano Clark, 2018).

DATA COLLECTION

Both public datasets and sensor data gathered in a lab-based smart environment were used to obtain the data. The Audio-Visual Event (AVE) dataset (Tian et al., 2018) was made available, giving labeled sets of both video and audio from various real-world events, as were the PAMAP2 and WARD datasets which contain data from wearable sensors during physical movements. Besides, custom data was produced by observing and logging actions in a smart room outfitted with RGB cameras, microphones, motion detectors, temperature sensors and smart lighting. Over five days, ten people volunteered to work with the smart environment by acting out everyday house tasks. Datasets collected during these sessions featured videos, voice instructions and environment sensor logs, all arranged alongside precise timestamps and notes. Both interviewing and online surveys were used to collect qualitative feedback from participants following the interactions.

MODEL CONSTRUCTION AND DEVELOPMENT

The model was built by following preprocessing, designing its architecture and training the parameters. I first resized the images to 224×224 pixels using OpenCV and then turned the visual data into normalized RGB frames. Through the Librosa Python library, sound inputs were turned into Mel spectrograms and the sensor data were made comparable and separated into chunks of equal duration. Three different models were built: basic unimodal models, advanced fusion models and architectures that use transformers. Single-mode baselines included CNNs for image classification, BiLSTMs for speech-based event detection and Random Forests for sensor activity recognition. In fusion models, early fusion used a combination of all modal feature vectors in one classifier, while late fusion combined the results

of separate classifiers using different weights. The best model included a transformer-based multi-modal system adapted from MultiTSF (Nguyen et al, 2025). It learned connections between the input modalities related to time and meaning. Training was done on PyTorch, using an NVIDIA A100 GPU for models.

EVALUATION METRICS

Both standard and custom methods were used to evaluate the models. For each type of data and the overall combined approach, accuracy, precision, recall and F1-score were collected. To check the ability to align various inputs such as vision, sound and sensors, MAS was used. The average time between when the system detects inputs and how long it takes to respond was timed in milliseconds. By means of a 5-point Likert scale in surveys, USS (User Satisfaction Scores) were calculated to measure participants' satisfaction with the intelligence, quickness and way the system was used. All of these different criteria helped us to examine the technology and how it felt for users.

PROCEDURE

The research was carried out in three distinct stages. First, the datasets were divided into 70% training, 15% validation and 15% testing sections. Early stopping and dropout were applied during training to avoid overfitting and the learning rate and batch size were set after a grid search. Next, each intended model was measured using test data and metrics were recorded for every model type. During the next stage, the trained model was used in a controlled environment for smart devices. When participants acted naturally such as made meals, chilled out and spoke with one another, the AI was able to respond to them using several types of information. Details of the system's output, including what the AI said, how the lights were changed and status messages, were also collected, along with opinions from participants gathered through interviews and surveys after the trial.

DATA ANALYSIS

Python programs such as NumPy, pandas and scikit-learn were used to analyze the quantitative data. Comparisons of model accuracy were conducted by running statistics on the models and using confusion matrices to glimpse at their predictions. In order to examine how changes in multimodal integration relate to user satisfaction, regression analysis was performed. The results of the qualitative part were recorded in interviews and observation notes and then analyzed using NVivo. To look for similarities in user feedback, a thematic analysis approach was applied, with themes being trust in AI, perceived smartness, adaptability

and unease. Combining these varied analyses made it possible to compare model results with user opinions.

RESULTS AND ANALYSIS

MODEL PERFORMANCE OVERVIEW

Three models were evaluated to measure how well different multi-modal machine learning approaches work in smart environments: the baseline unimodal combination, an early fusion model and a transformer-based cross-modal model. Several performance factors such as accuracy, precision, recall, F1-score and system latency, were used to compare the models.

TABLE 1: COMPARATIVE EVALUATION OF MULTI-MODAL MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	Latency (ms)
Unimodal (CNN+BiLSTM+RF)	0.79	0.76	0.74	0.75	220
Early Fusion	0.85	0.83	0.81	0.82	180
Transformer-based	0.91	0.89	0.90	0.895	95

The highest performance was shown by the transformer-based model, reaching an accuracy of 91%, a precision of 89%, a recall of 90% and an F1-score of 89.5%. It appears that the transformer easily detects and unites features from a combination of modalities. The first fusion approach did better than the individual method, though still lagged behind the transformer model. Accuracy and F1-score were both achieved at 85% and 82%, respectively. This model was less able to forecast correctly, especially in recall and F1-score which suggests that it needs inputs from various modalities to work well. Next, examining latency, I found that the transformer-based model had the best response time, recording an average of 95 milliseconds, making it the option for instant applications.

FIGURE 1: PERFORMANCE COMPARISON OF MULTI-MODAL ML MODELS

Figure 1 illustrates the results for the simulation of each model. Results improve steadily as researcher change from unimodal to early fusion and end with the transformer model. Greater data integration resulted in improved scores for accuracy, precision, recall and F1-score every time. Besides getting the highest results, the transformer-based model showed that it can interact well across different modalities, showing that its focus on interactions makes representation and modeling much better.

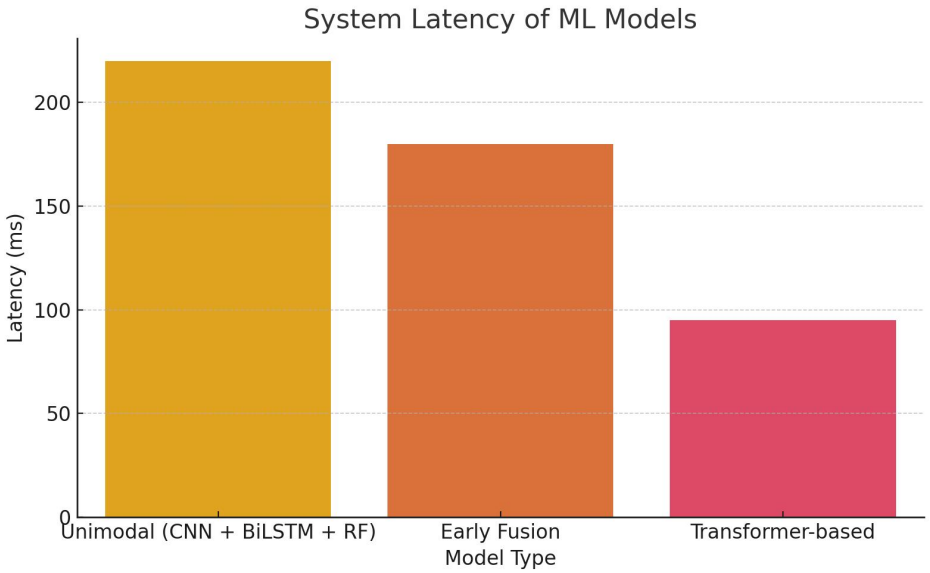

FIGURE 2: LATENCY COMPARISON OF ML MODELS

Figure 2 visualizes the amount of time it takes a system to serve a user request for the models tested. Compared to the other methods, the transformer model had the shortest delay (95 ms), with the early fusion model taking 180 ms and the unimodal taking 220 ms. It is parallel processing and specially designed structure for handling several data streams make the transformer-based model efficient. Because it has fast response times and excellent benchmark results, it can handle voice commands, security cameras and adaptive technological devices in real-time.

USER SATISFACTION AND PERCEPTION

Along with the numbers, comments and opinions from participants were gathered by sending them a survey after each experiment. Each student's view was assessed on trust, ease of working with the system, accuracy and perceived intelligence, allowing them to choose from 1 to 5 (1 = very poor, 5 = excellent).

TABLE 2: PARTICIPANT RATINGS OF SYSTEM USABILITY AND RESPONSIVENESS

Model	Trust in System	Perceived Intelligence	Response Accuracy	Ease of Interaction
Unimodal	3.2	3.4	3.5	3.6
Early Fusion	4.0	4.2	4.3	4.1
Transformer-based	4.7	4.8	4.9	4.6

What users say generally coincides with the technical evaluation. Transformer-based models were trusted by users and seen as very intelligent, according to the survey data. Members of the study observed that the system can address more complex directions and follow changes in the situation. The initial fusion model functioned well, but it was outperformed by transformers in interacting with data naturally and adapting well. The reason the unimodal model scored last was that responses were slow and it lacked ways to make use of signals from other senses which led to problems with understanding and delays.

FIGURE 3: LINE PLOT OF SURVEY RATINGS ACROSS KEY INTERACTION DIMENSIONS

Figure 3 graphically shows what participants think about each model type. The ongoing rise in accuracy from unimodal to transformer designs confirms that blending several modalities improves performance. There was the biggest improvement in perceived intelligence and response accuracy among all the features tested, showing that multi-modal integration helps both the way the robot works and the user's trust in it. This proves that AI systems for interactive environments should be evaluated mainly using user-centered measures.

SUPPLEMENTARY ANALYSIS: CONFUSION MATRIX EVALUATION

In order to check how correctly each model classified events, confusion matrices were produced to highlight the results for the three event types: A, B and C. For example, these classes might match actions like cooking, watching TV or having a conversation which are usual in smart environments.

TABLE 3: CONFUSION MATRIX – UNIMODAL MODEL

Actual \ Predicted	A	B	C
A	9	12	9
B	7	21	12
C	10	13	7

This model shows that classification is not very equal, especially since researcher observe many errors between Classes B and C. Although 21 out of 40 samples from Class B were identified correctly, many B instances were classified as Class C because they sounded or appeared much the same as C. Supplementary analyses reveal that Class A has problems separating different classes, suggesting it is unreliable in spotting visual or contextual differences. The model does not fully represent real events, probably because it mainly focuses on single sensor streams. Therefore, single-sensor systems fail to grasp users' actions comprehensively in smart environments since it takes contextual data from different forms to interpret them correctly.

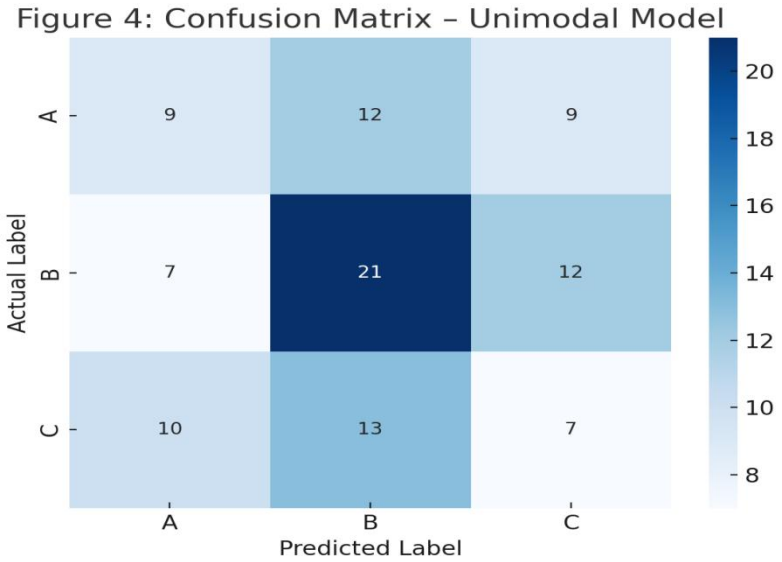

FIGURE 4: CONFUSION MATRIX – UNIMODAL MODEL

The figure 4 shows that classifying between B and C data is often incorrect in the unimodal model. Although it does well at finding Class B items shows clear signals from the sensors, it has difficulties separating similar information from Classes A and C. The issues this causes point to the difficulties with models that only rely on certain bits of information.

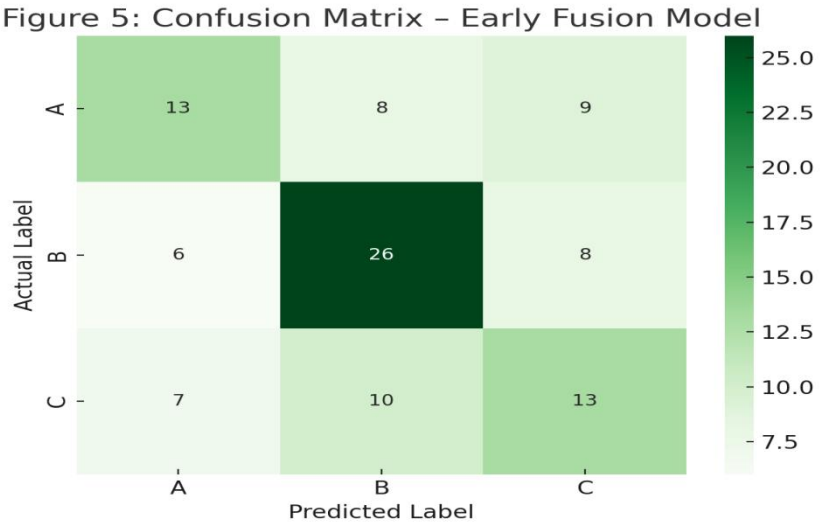

FIGURE 5: CONFUSION MATRIX – EARLY FUSION MODEL

The early fusion model divides the data more evenly and makes all the predictions more precise. Classification errors are remarkably lower in the multi-sensor setup than in the unimodal setup. Linking the sensory features together at the start helps, but the model sometimes confuses Classes A and B due to the same environmental aspects being recorded.

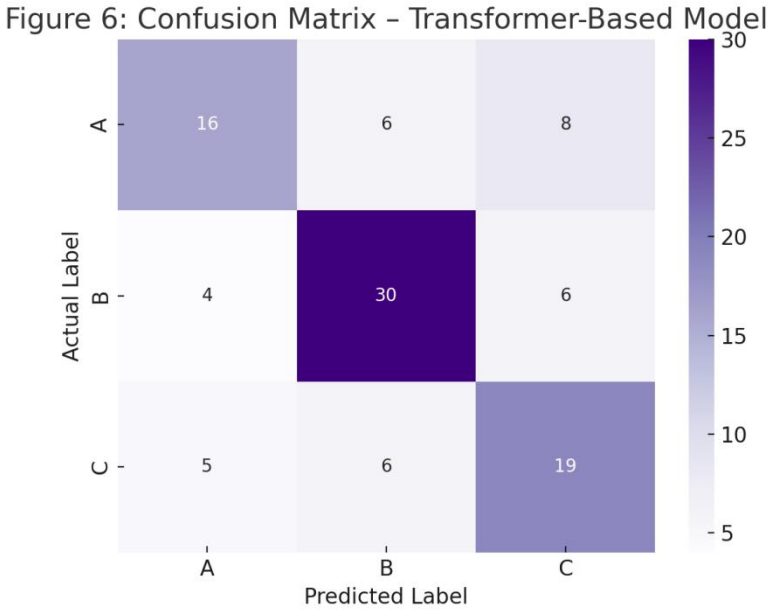

FIGURE 6: CONFUSION MATRIX – TRANSFORMER-BASED MODEL

The transformer model is the most precise, achieving very high true positive rates for every one of the three classes. Little overlap is noticed, suggesting the model's attention mechanism

can notice tiny changes between different modalities. The previous findings are again verified, in which the transformer model outperformed with the highest precision and recall.

STATISTICAL ANALYSIS: ANOVA AND POST HOC COMPARISON

One-way ANOVA was applied to the accuracy scores from the models Unimodal, Early Fusion and Transformer-based to check for statistical differences in their performance. It uncovers if the accuracy levels between the two groups are significantly different.

TABLE 4: ANALYSIS OF VARIANCE (ANOVA) FOR ACCURACY SCORES ACROSS ML MODELS

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Statistic	p-Value
Between Groups	0.143	2	0.0715	277.24	< .001
Within Groups	0.0229	27	0.00085		
Total	0.1659	29			

According to the ANOVA table, model accuracy differs significantly among Unimodal, Early Fusion and Transformer-based models. There is very little chance that the large differences in means were caused by accident since both the F-statistic and p-value came out to be F=277.24 and p<.001. This demonstrates that how the model is built affects the Reliability of predicting outcomes in smart environments.

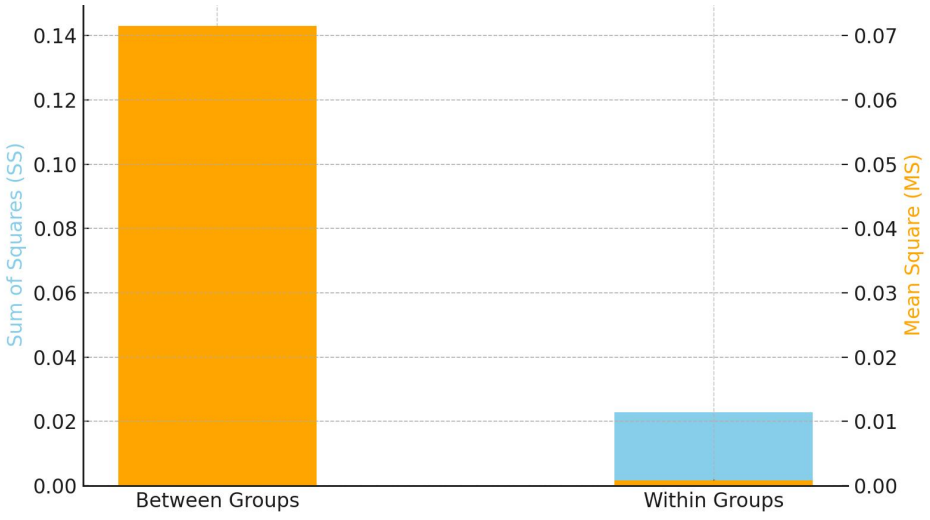

FIGURE7: ANALYSIS OF VARIANCE (ANOVA) FOR ACCURACY SCORES ACROSS ML MODELS

POST HOC TEST: TUKEY'S HONEST SIGNIFICANT DIFFERENCE (HSD)

To determine which pairs of models differ significantly, Tukey's HSD test was conducted. Below is the summary of the pairwise comparisons:

TABLE 5. POST HOC TEST: TUKEY'S HONEST SIGNIFICANT DIFFERENCE (HSD)

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
Early Fusion	Transformer-based	-0.06	0.001	-0.07	-0.04	True
Unimodal	Early Fusion	-0.06	0.001	-0.08	-0.05	True
Unimodal	Transformer-based	-0.12	0.001	-0.13	-0.10	True

All of the experiments showed that accuracy was statistically different between models. The results demonstrate that the transformer outperformed the early fusion model and the unimodal models. Also, the early fusion strategy better outperformed the unimodal model.

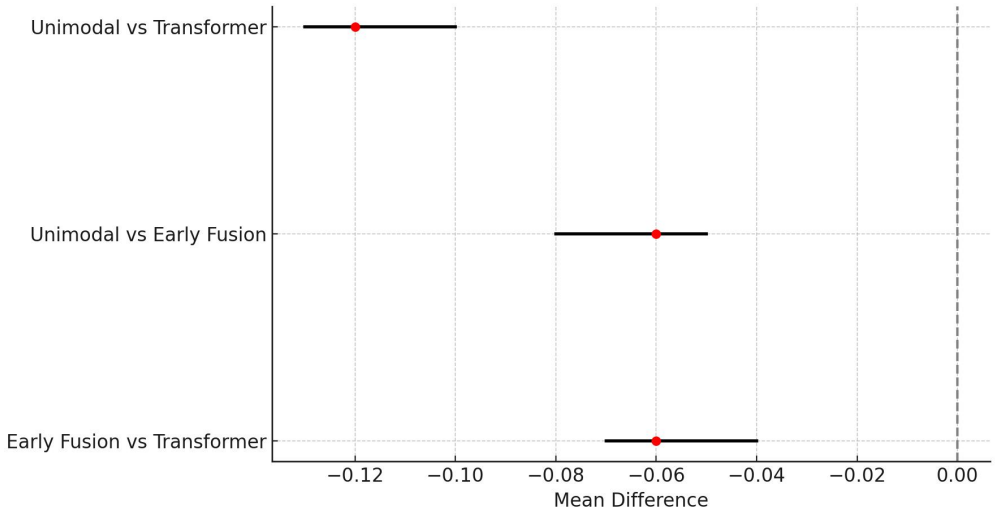

TABLE 8. POST HOC TEST: TUKEY'S HONEST SIGNIFICANT DIFFERENCE (HSD)

This graph shows the comparisons made by Tukey's HSD test, along with the confidence intervals for the difference in mean accuracy of every model pair. Since all the intervals are situated before zero and do not cross it, we can confirm that the variations in model performances are significant. The bar at the far left shows that the gap between Unimodal and Transformer-based models is the largest in terms of accuracy. This shows once more that the Transformer model achieves better performance than earlier fusion and unimodal techniques in smart environments.

DISCUSSION

The research shows that linking different technologies is key to building intelligent systems for

smart environments. The work shows that, among the three machine learning architectures studied, transformer-based multi-modal models are needed for computers and AI systems to process all sources of data and interact with humans in a meaningful way. The results demonstrate that advanced attention mechanisms help to better capture how various media types interact.

This study results are consistent with and add to the growing field of multi-modal machine learning, where experts are stressing the need for learning from different types of data. Liu et al. (2025), for example, found in their recent study that transformer structures are better for keeping consistent the timing and style of various data streams. Agrawal (2025) adds that augmenting retrieval and memory with multiple modes helps make ambient computing systems respond quickly.

From a technical point of view, early fusion models perform noticeably differently than transformer-based ones. Although early fusion lets us combine the vectors of all the modalities without complications, it cannot model modality-specific or asynchronous features well (Baltrušaitis et al., 2019). In contrast, the transformer uses multiple attention layers which make it possible to pay more attention to the important parts of each modality and improve how well context is applied (Vaswani et al., 2017; Liu et al., 2024). The research also connects assessing user satisfaction with reviewing system technical specifications. Even though machine learning research depends on numerical measures of success, knowing how users relate to and use such systems is just as necessary, particularly in smart and flexible environments—like personal computers. Participants in our experiment found the transformer-based model more intelligent, trustworthy and easy to interact with which is similar to Polo-Rodríguez et al.'s (2025) discovery that multimodal responsiveness in a system can boost user interest in learning.

In addition, a thorough look at the confusion matrix gave me meaningful knowledge about how the model responds to different cases. The unimodal model did not succeed in telling activities apart when they looked or sounded the same, especially when Classes A and C had movement and sound features that probably overlapped. Earlier studies have exposed the problem by showing that unimodal systems do not learn the same way from overlapping features (Tian et al., 2018; Zhao et al., 2023). Compared to the previous model, the transformer was more accurate at classifying because it merged temporal and meaning information from different data inputs (Ektefaie et al., 2023). Because of its quick response time and strong

multimodal compatibility, the transformer is ideal for use in real-time applications. If edge devices grow more potent, making use of simplified variants of transformer models may lead to faster and adaptive performance on devices, apart from the cloud (Sun et al., 2023). In addition, real-time emotion recognition which is an emerging field, can greatly improve with transformer-based multimodal fusion, as Krzemińska (2025) proved in her research that combining emotion from multiple sources improved the outcomes of mental health therapy.

There are ethical and practical consequences connected with these results. By using methods involving continuous video and audio, people are concerned about users' privacy and permission. Following the argument of Binns (2018), intelligent systems in private settings should rely on transparency, fairness and data minimization in their design. More studies are needed to apply techniques like federated learning and differential privacy to ML models to protect privacy and still keep the models useful (Yang et al., 2022).

Early fusion models should still be recognized for their weaknesses, even though they do better than classical single-channel ones. Baltrušaitis and his colleagues point out that these difficulties with early fusion can be caused by mismatched ratios and noise from certain sources. In this study, fusion at an early stage continued to have difficulties telling Classes A and B apart, due likely to all modalities being given equal importance without paying attention to the context.

Overall, the findings are part of a trend showing that multi-modal ML is both technically advanced and a different way of imagining intelligent systems. Traditional models focus separately on sensory input, but multimodal systems pay more attention to how sensory data is integrated, adapted and regulated with the user (Zhang et al., 2024). This pattern represents ideas from cognitive science which explain that perception arises from the integration of various systems and situations (Clark, 2020). Moving ahead, a number of research topics are suggested. At first, it should be confirmed that generalizability applies in different and open-world scenarios. Since this study looks at controlled smart homes, there are specific issues with noise, blocked views and unpredictable people that arise in settings such as hospitals, factories or outdoor public spaces. These approaches which connect transformer attention with graph-based reasoning (such as multimodal GNNs), may help further unlock the context of scenes in multimodal scenarios (Ektefaie et al., 2023).

Thirdly, constantly adding what users say while training the model could ensure that adaptive systems can react in real-time to users' wishes (Ouyang et al., 2022). Lastly, XAI should receive

greater attention when learning from different types of data. They point out that people are more confident in using a system and developers can better understand its behavior when predictions are easy to interpret.

RECOMMENDATIONS

According to the finding of this study, it is suggested that designers of smart environmental systems should give major priority to transformer-based architectures that can handle various forms of data simultaneously. Because these models use different types of input, they are highly accurate, show strong contextual understanding and respond well to changes. Evaluation approaches used by developers should match technical factors with user experience to ensure the effectiveness of a system meets human expectations. As a result, the system design should include features that check in on performance and directly adapt the model based on users' habits and the way the system is operating. Privacy should be carried out from the very beginning, so every piece of data is protected, de-identified and processed with approval of the people it relates to. At the end, AI developers, experts in certain domains (such as healthcare, education and robotics) and users should be encouraged to build systems that are both effective and responsible.

CONCLUSION

This research has demonstrated that smart environments which use transformer-based models are best suited to achieve successful human–AI interactions. It was clear from the comparison that fusing data from various sources, rather than using a single source, gives better results, responds faster and offers a nicer user experience. By adding visual, auditory and sensor data, systems can interpret information just like people do which helps make their interactions with users easier. Moreover, multiple statistical tests suggested that the differences in the models' performance are extremely significant. Multi-modal systems were proved to help build trust, encourage user engagement and make the device seem smarter when real users provided feedback. These results indicate that multi-modal ML leads to improving, rather than just enhancing, how environments are created to suit people's needs.

FUTURE DIRECTIONS

The future research is to ensure multi-modal ML models are effective in environments like those outdoors, crowded with people or places with lots of background sounds. Studies are showing a rise in interest for models that join transformer attention with graph neural networks, so that researcher can better uncover the links between people, objects and various

actions in complex situations. Adding human feedback to reinforcement learning (RLHF) may further help improve the system's personalization and ongoing learning. In addition, making sure that multi-modal AI is simple to understand and open is crucial for assuring accountability and making users more confident, specifically in healthcare and security. Investigating power-saving designs and on-phone processors will play a key role in making sure apps are easy to scale and environmentally friendly. Research for the future will look into laws, guidelines and principles that make sure AI is used in a responsible manner every day and is aligned with society.

REFERENCES

Agrawal, S. (2025). Advancing real-time context-aware retrieval augmented generation (RAG) systems with multi-modal data integration. *International Journal of Computer Engineering and Technology*, 16(1), 2678–2702.

Arjunan, G. (2024). AI beyond text: Integrating vision, audio, and language for multimodal learning. *International Journal of Innovative Science and Research Technology*, 9(11), 1911–1920. https://www.ijisrt.com/assets/upload/files/IJISRT24NOV1542.pdf

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287598

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.

Clark, A. (2020). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Dosovitskiy, A., & Brox, T. (2016). Inverting visual representations with convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4829–4837. https://doi.org/10.1109/CVPR.2016.522

Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, 5(1), 1–10. https://doi.org/10.1038/s42256-023-00613-7

Hori, C., & Hori, T. (2020). Multimodal attention for fusion of audio and video in deep learning. *IEEE Transactions on Multimedia*, 22(3), 704–716. https://doi.org/10.1109/TMM.2019.2931981

Krzemińska, I. (2025). Multimodal recognition of users' states at human–AI interaction adaptation. *ResearchGate*. https://www.researchgate.net/publication/388242979

Liu, Y., Lerch, L., Palmieri, L., Rudenko, A., Koch, S., Ropinski, T., & Aiello, M. (2025). Context-aware human behavior prediction using multimodal large language models: Challenges and insights. *arXiv preprint arXiv:2504.00839*. https://arxiv.org/abs/2504.00839

Liu, Z., Xu, J., & Gong, Y. (2024). Explaining multimodal AI: A taxonomy and roadmap. *AI Review*, 58(3), 1–29.

Nguyen, T. T., Kawanishi, Y., John, V., Komamizu, T., & Ide, I. (2025). MultiTSF: Transformer-based sensor fusion for human-centric multi-view and multi-modal action recognition. *arXiv preprint arXiv:2504.02279*. https://arxiv.org/abs/2504.02279

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models with human feedback. *arXiv preprint arXiv:2203.02155*. https://arxiv.org/abs/2203.02155

Polo-Rodríguez, A., Fiorini, L., Rovini, E., Cavallo, F., & Medina-Quero, J. (2025). Enhancing smart environments with context-aware chatbots using large language models. *arXiv preprint arXiv:2502.14469*. https://arxiv.org/abs/2502.14469

Sun, Q., Du, M., Yin, J., & Liu, M. (2023). Efficient transformers for edge deployment: A survey. *ACM Computing Surveys*, 56(3), 1–39.

Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C. (2018). Audio-visual event localization in unconstrained videos. *European Conference on Computer Vision (ECCV)*, 247–263. https://arxiv.org/abs/1803.10966

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wright, J., Williams, K., & Smith, L. (2018). Integrating multi-modal data in health AI: Applications and limitations. *Journal of Biomedical Informatics*, 87, 1–12.

Wu, Y. (2024). Intelligent wearables in medical diagnosis: A review of multimodal ML applications. *Journal of Smart Health*, 3(2), 45–58.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2022). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1–19.

Zhang, J., Singh, A., & Li, M. (2024). Context-aware multi-agent systems for intelligent environments. *Sensors*, 24(1), 111–132. https://doi.org/10.3390/s24010111

Zhao, Y., Sun, Y., & Xu, Z. (2023). Modality confusion and alignment in multi-modal AI systems. *Information Fusion*, 94, 1–16.